

Some Classical and Some New Ideas for Identification of Linear Systems [★]

Lennart Ljung,

Division of Automatic Control, Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden

Abstract

This paper gives an overview of identification of linear systems. It covers the classical approach of parametric methods using Maximum Likelihood and Prediction Error Methods, as well all classical non-parametric methods through spectral analysis.

It also covers very recent techniques dealing with convex formulations by regularization of FIR and ARX models, as well as new alternatives to spectral analysis, through local linear models.

An example of identification of aircraft dynamics illustrates the approaches.

Key words: System identification; Maximum Likelihood, Prediction Error Methods, Spectral Analysis, Regularization, Local Polynomial Methods

1 Introduction

1.1 An Introductory Example: Aircraft Dynamics

Consider a physical system, with observed input and output signals, see Figure 1. Let us take a modern military aircraft, like the Swedish fighter Gripen, as an example. From one of the earlier test flights, some data



Fig. 1. The Swedish aircraft Gripen

were recorded as depicted in Figure 2.

[★] Support from the European Research Council under the advanced grant LEARN, contract 267381 is gratefully acknowledged.

Email address: ljung@isy.liu.se (Lennart Ljung).

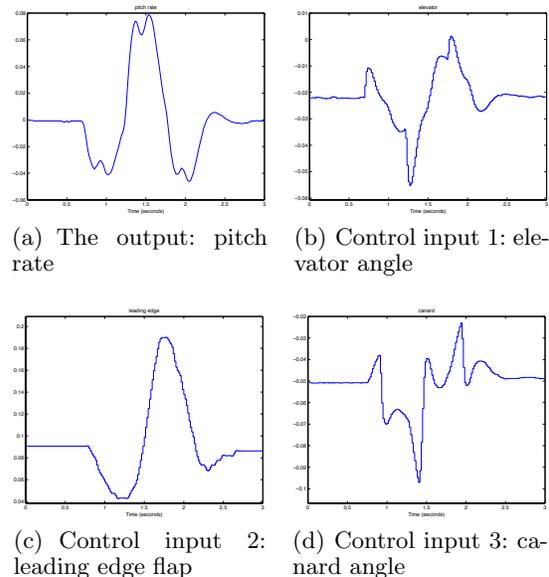


Fig. 2. Data from an early test flight of Gripen. These data cover 3 seconds of flight and are sampled at 60 Hz.

In order to be able to simulate the aircraft and to design an effective autopilot, it is necessary to understand how, in this case, the pitch rate is affected by the three inputs. We need mathematical expressions for this. A fair amount of knowledge exists about aircraft dynamics, but let us just try a simple difference equation relation. Denote the output, the pitch rate, at sample num-

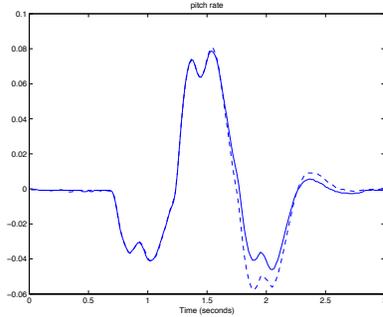


Fig. 3. The measured output (solid line) compared to the 5 step ahead predicted one (dashed line).

ber t by $y(t)$ and three control inputs at the same time by $u_k(t)$, $k = 1, 2, 3$. Then assume that we can write

$$\begin{aligned}
 y(t) = & a_1y(t-1) - a_2y(t-2) - a_3y(t-3) \\
 & + b_{1,1}u_1(t-1) + b_{1,2}u_1(t-2) \\
 & + b_{2,1}u_2(t-1) + b_{2,2}u_2(t-2) \\
 & + b_{3,1}u_3(t-1) + b_{3,2}u_3(t-2)
 \end{aligned} \tag{1}$$

In this simple relationship we can adjust the parameters to fit the observed data as well as possible by a common Least Squares fit. We use only the 90 first data points of the observed data. That gives certain numerical values of the 9 parameters above:

$$\begin{aligned}
 y(t) - 1.15y(t-1) + 0.50y(t-2) - 0.35y(t-3) \\
 = -0.54u_1(t-1) + 0.04u_1(t-2) \\
 + 0.15u_2(t-1) + 0.16u_2(t-2) \\
 + 0.16u_3(t-1) + 0.07u_3(t-2)
 \end{aligned} \tag{2}$$

We may note that this model is unstable – it has a pole in 1.0026, but that is in order, because the pitch channel is unstable at the velocity and altitude in question.

How can we test if this model is OK? Since we used only half of the observed data for the estimation we can test the model on the whole data record. Since the model is unstable and thus simulation is difficult, it is natural to let the model predict future outputs, say 5 samples ahead, and compare with the measured outputs. That is done in Figure 3. We see that the simple model (2) provides quite reasonable predictions over data it has not seen before. The could conceivably be improved if more elaborate model structures than (1) were tried out.

1.2 System Identification

System Identification is about building mathematical models of dynamical systems from observed input-output signals, like we did in (2). This problem area contains a number of considerations, like

- what model type, e.g. (1) should be used?

- how should the parameters in the model be adjusted?
- what inputs should be applied to obtain a good model?
- how do we assess the quality of the model?
- how do we gain confidence in an estimated model?
-

There is a very extensive literature on the subject, with many text books, like [3] and [12]. Most of the techniques for system identification have their origins in estimation paradigms from mathematical statistics, and classical methods like Maximum Likelihood (ML) have been important elements in the area. In this article the main ingredients of this “classical” view of System Identification will be reviewed. Quite recently, alternative techniques, mostly from machine learning and convex optimizations, but also with the roots from classical statistics have emerged. The main elements of these will also be reviewed here.

2 Classical Approach to Parametric Methods

2.1 Model Structures

A model structure \mathcal{M} is a parameterized collection of models that describe the relations between the input and output signal of the system. The parameters are denoted by θ so $\mathcal{M}(\theta)$ is a particular model. That model gives a rule to predict (one-step-ahead) the output at time t , i.e. $y(t)$, based on observations of previous input-output data up to time $t-1$ (denoted by Z^{t-1}).

$$\hat{y}(t|\theta) = g(t, \theta, Z^{t-1}) \tag{3}$$

For linear systems, a general model structure is given by the transfer function G from input to output and the transfer function H from a white noise source e to output additive disturbances:

$$y(t) = G(q, \theta)u(t) + H(q, \theta)e(t) \tag{4a}$$

$$Ee^2(t) = \lambda; \quad Ee(t)e(k) = 0 \text{ if } k \neq t \tag{4b}$$

This model is in discrete time and q denotes the shift operator $qy(t) = y(t+1)$. We assume for simplicity that the sampling interval is one time unit. The expansion of $G(q, \theta)$ in the inverse (backwards) shift operator gives the *impulse response* of the system:

$$G(q, \theta) = \sum_{k=1}^{\infty} g_k(\theta)q^{-k}u(t) = \sum_{k=1}^{\infty} g_k(\theta)u(t-k) \tag{5}$$

The discrete time Fourier Transform, gives the *frequency response* of the system:

$$G(e^{i\omega}, \theta) = \sum_{k=1}^{\infty} g_k(\theta)e^{ik\omega} \tag{6}$$

The natural predictor for(4a) is

$$\hat{y}(t|\theta) = \frac{H(q, \theta) - 1}{H(q)}y(t) + \frac{G(q, \theta)}{H(q, \theta)}u(t) \quad (7)$$

Note that the expansion of H starts with a "1", so the numerator in the first term starts with h_1q^{-1} so there is a delay in y . The question now is how to parameterize G and H .

2.1.1 Black-Box models

Common *black box* (i.e. no physical insight or interpretation) parameterizations are to let G and H be rational in the shift operator:

$$G(q, \theta) = \frac{B(q)}{F(q)}; \quad H(q, \theta) = \frac{C(q)}{D(q)} \quad (8a)$$

$$B(q) = b_1q^{-1} + b_2q^{-2} + \dots + b_nq^{-nb} \quad (8b)$$

$$F(q) = 1 + f_1q^{-1} + \dots + f_nq^{-nf} \quad (8c)$$

$$\theta = [b_1, b_2, \dots, f_n] \quad (8d)$$

C and D are like F *monic*, i.e. start with a "1".

A very common case is that $F = D = A$ and $C = 1$ which gives the *ARX-model*:

$$y(t) = \frac{B(q)}{A(q)}u(t) + \frac{1}{A(q)}e(t) \text{ or} \quad (9a)$$

$$A(q)y(t) = B(q)u(t) + e(t) \text{ or} \quad (9b)$$

$$y(t) + a_1y(t-1) + \dots + a_nay(t-na) \quad (9c)$$

$$= b_1u(t-1) + \dots + b_nu(t-nb) \quad (9d)$$

This is the model structure we used in (1) in the introductory example.

Other common black/box structures of this kind are FIR (Finite Impulse Response model, $F = C = D = 1$), ARMAX ($F = D = A$), and BJ (Box-Jenkins, all four polynomial different.)

2.1.2 Grey-Box Models

If some physical facts are known about the system, it is possible to build in that into a *Grey-Box Model*. It could, for example be that for the airplane in the introduction, the motion equations are known from Newton's laws, but certain parameters are unknown, like the aerodynamical derivatives. Then it is natural to build a continuous time state-space models from physical equations:

$$\begin{aligned} \dot{x}(t) &= A(\theta)x(t) + B(\theta)u(t) \\ y(t) &= C(\theta)x(t) + D(\theta)u(t) + v(t) \end{aligned} \quad (10)$$

Here θ corresponds to unknown physical parameters, while the other matrix entries signify known physical behaviour. This model can be sampled with the well-known sampling formulas to give

$$\begin{aligned} x(t+1) &= \mathcal{F}(\theta)x(t) + \mathcal{G}(\theta)u(t) \\ y(t) &= C(\theta)x(t) + D(\theta)u(t) + w(t) \end{aligned} \quad (11)$$

See [8] for deeper discussion of sampling of systems with disturbances.

The model (11) has the transfer function from u to y

$$G(q, \theta) = C(\theta)[qI - \mathcal{F}(\theta)]^{-1}\mathcal{G}(\theta) + D(\theta) \quad (12)$$

so we have achieved a particular parameterization of the general linear model (4a).

2.2 Fitting Time-Domain Data

Suppose now we have collected a data record in the time domain

$$Z^N = \{u(1), y(1), \dots, u(N), y(N)\} \quad (13)$$

It is the most natural to compare the model predicted values (7) with the actual outputs and form the criterion of fit

$$V_N(\theta) = \frac{1}{N} \sum_{t=1}^N [y(t) - \hat{y}(t|\theta)]^2 \quad (14)$$

and form the parameter estimate

$$\hat{\theta}_N = \arg \min V_N(\theta) \quad (15)$$

We call this the *Prediction Error Method, PEM*. It coincides with the *Maximum Likelihood, ML*, method if the noise source e is Gaussian. See, e.g. [3] or [6] for more details.

2.3 Fitting Frequency-Domain Data

Suppose instead that we have been given frequency domain data. That could be in the input-output form

$$Y_N(e^{i\omega_k}), U_N(e^{i\omega_k}), k = 1, 2, \dots, M \quad (16)$$

$$Y_N(z) = \frac{1}{\sqrt{N}} \sum_{k=1}^N y(k)z^{-k} \quad (17)$$

or being observed samples from the frequency function

$$\hat{G}_N(e^{i\omega_k}), k = 1, 2, \dots, M \quad (18)$$

$$\text{e.g. } \hat{G}_N(e^{i\omega}) = \frac{Y_N(e^{i\omega})}{U_N(e^{i\omega})} \text{ (ETFE)} \quad (19)$$

By taking the Fourier transform of (4a) we see that

$$Y(e^{i\omega}) = G(e^{i\omega}, \theta)U(e^{i\omega}) \quad (20)$$

plus a noise term that has variance

$$n = \lambda |H(e^{i\omega}, \theta)|^2 \quad (21)$$

Simple least squares (LS) curve fitting says that we should fit observations with weights that are inversely proportional to the measurement variance. That gives the weighed LS criterion

$$V_N(\theta) = \sum_{k=1}^M |Y(e^{i\omega_k}) - G(e^{i\omega_k}, \theta)U_N(e^{i\omega_k})|^2 / |H(e^{i\omega_k}, \theta)|^2 \quad (22)$$

(the constant λ does not effect the minimization of V_N).

It can readily be verified that (22) coincides with (14) by Parseval's identity in case $M = N$ and the frequencies ω_k are selected as the DFT grid.

Notice that (22) can be written as

$$V_N(\theta) = \sum_{k=1}^M \left| \frac{Y_N(e^{i\omega_k})}{U_N(e^{i\omega_k})} - G(e^{i\omega_k}, \theta) \right|^2 \cdot \left| \frac{U_N(e^{i\omega_k})}{H(e^{i\omega_k}, \theta)} \right|^2 \quad (23)$$

We can see that as a properly weighted curve-fitting of the frequency function to the ETFE (19).

3 Bias and Variance

The observations, certainly of the output from the system are affected by noise and disturbances, which of course also will influence the estimated model (15). The disturbances are typically described as stochastic processes, which makes the estimate $\hat{\theta}_N$ a *random variable*. This has a certain probability distribution function (pdf) and a mean and a variance. The difference between the mean and a true description of the system measures the *bias* of the model. If the mean coincides with the true system, the estimated is said to be *unbiased*. The total error in a model thus has two contributions: the bias and the variance.

3.1 Trade-off between bias and variance

Generally speaking the quality of the model depends on the quality of the measured data and the flexibility of the chosen model structure (3). A more flexible model structure typically has smaller bias, since it is easier to come closer to the true system. At the same time, it will

have a higher variance: With higher flexibility it is easier to be fooled by disturbances. So the trade-off between bias and variance to reach a small total error is a choice of balanced flexibility of the model structure.

As the model gets more flexible, the fit to the estimation data in (15), $V_N(\hat{\theta}_N)$ will always improve. To account for the variance contribution, it is thus necessary to modify this fit to assess the total quality of the model. A much used technique for this is Akaike's criterion, e.g. [1],

$$\hat{\theta}_N = \arg \min \left[V_N(\theta) + 2 \frac{\dim \theta}{N} \right] \quad (24)$$

were the minimization also take place over a family of model structures with different number of parameters ($\dim \theta$).

Another important technique is to evaluate the criterion function for the model for another set of data, *validation data*, and pick the model which gives the best fit to this independent data set. This is known as *cross validation*.

3.2 Asymptotic Properties of the Model

Except in simple special cases it is quite difficult to compute the pdf of the estimate $\hat{\theta}_N$. However, its *asymptotic properties* as $N \rightarrow \infty$ are easier to establish. The basic results can be summarized as follows: (\mathcal{E} denotes mathematical expectation)

-

$$\hat{\theta}_N \rightarrow \theta^* = \arg \min_{\theta} \mathcal{E} \lim_{N \rightarrow \infty} V_N(\theta) \quad (25)$$

So the estimate will converge to the best possible model, which gives the smallest average prediction error.

-

$$\text{Cov} \hat{\theta}_N \sim \frac{\lambda}{N} \left[\text{Cov} \frac{d}{d\theta} \hat{y}(t|\theta) \right]^{-1} \quad (26)$$

So the covariance matrix of the parameter estimate is given by the inverse covariance matrix of the gradient of the predictor wrt the parameters. λ is the variance of the optimal prediction errors (*the innovations*). See [3], chapters 8 and 9 for a general treatment.

These results are valid for quite general model structures. Now, specialize to linear models (4a) and assume that the true system is described by

$$y(t) = G_0(q)u(t) + H_0(q)e(t) \quad (27)$$

which could be general transfer functions, possibly much more complicated than the model. Then

•

$$\theta^* = \arg \min_{\theta} \int_{-\pi}^{\pi} |G(e^{i\omega}, \theta) - G_0(e^{i\omega})|^2 \frac{\Phi_u(\omega)}{|H(e^{i\omega}, \theta)|^2} d\omega \quad (28)$$

That is, the frequency function of the limiting model will approximate the true frequency function as well as possible in a frequency norm given by the input spectrum Φ_u and the noise model.

•

$$\text{Cov}G(e^{i\omega}, \hat{\theta}_N) \sim \frac{n}{N} \frac{\Phi_v(\omega)}{\Phi_u(\omega)} \text{ as } n, N \rightarrow \infty \quad (29)$$

where n is the model order and Φ_v is the noise spectrum $\lambda|H_0(e^{i\omega})|^2$. The variance of the estimated frequency function at a given frequency is thus, for a high order model proportional to the Noise-to-Signal ratio at that frequency. That is a natural and intuitive result.

4 Approximating Linear Systems by ARX Models

Suppose the true linear system is given by

$$y(t) = G_0(q)u(t) + H_0(q)e(t) \quad (30)$$

Suppose we build an ARX model (9) for larger and larger orders $n = na$ and $m = nb$:

$$A_n(q)y(t) = B_m(q)u(t) + e(t) \quad (31)$$

Then it is well known from [5] that as the orders tend to infinity at the same time as then number of data N increases even faster we have for the ARX estimate

$$\frac{\hat{B}_m(q)}{\hat{A}_n(q)} \rightarrow G_0(q) \quad (32a)$$

$$\frac{1}{\hat{A}_n(q)} \rightarrow H_0(q) \text{ as } n, m \rightarrow \infty \quad (32b)$$

This is quite a useful result. ARX-models are easy to estimate. The estimates are calculated by linear least squares techniques, which are convex and numerically robust. Estimating a high order ARX model, possibly followed by some model order reduction could thus be a viable alternative to the numerically more demanding general PEM criterion (15). This has been extensively used, e.g. by [14], [15].

The only drawback with high order ARX-models is that they may suffer from high variance. That is the problem we now turn to.

5 Regularization of Linear Regression Models

5.1 Linear Regressions

A *Linear Regression problem* has the form

$$y(t) = \varphi^T(t)\theta + e(t) \quad (33)$$

Here y (the output) and φ (the regression vector) are observed variables, e is a noise disturbance and θ is the unknown parameter vector. In general $e(t)$ is assumed to be independent of $\varphi(t)$.

It is convenient to rewrite (33) in vector form, by stacking all the elements (rows) in $y(t)$ and $\varphi^T(t)$ to form the vectors (matrices) Y and Φ and obtain

$$Y = \Phi\theta + E \quad (34)$$

The least squares estimate of the parameter θ is

$$\hat{\theta}_N = \arg \min |Y - \Phi\theta|^2 \text{ or} \quad (35a)$$

$$\hat{\theta}_N = R_N^{-1}F_N; R_N = \Phi^T\Phi; F_N = \Phi^TY \quad (35b)$$

5.2 Regularized Least Squares

It can be shown that the variance of $\hat{\theta}$ could be quite large, in particular if Φ has many columns and/or is ill-conditioned. Therefore it makes sense to *regularize* the estimate by a matrix P :

$$\hat{\theta}_N = \arg \min |Y - \Phi\theta|^2 + \theta^T P^{-1}\theta \text{ or} \quad (36a)$$

$$\hat{\theta}_N = (R_N + P^{-1})^{-1}F_N; \quad (36b)$$

The presence of the matrix P will improve the numerical properties of the estimation and decrease the variance of the estimate, at the same time as some bias is introduced. Suppose that the data have been generated by (34) for a certain “true” vector θ_0 with noise with variance $\mathcal{E}EE^T = I$. (\mathcal{E} denotes mathematical expectation.) Then, the mean square error (MSE) of the estimate is

$$\mathcal{E}[(\hat{\theta}_N - \theta_0)(\hat{\theta}_N - \theta_0)^T] = (R_N + P^{-1})^{-1} \times (R_N + P^{-1}\theta_0\theta_0^T P^{-1})(R_N + P^{-1})^{-1} \quad (37)$$

A rational choice of P is one that makes this MSE matrix small. How shall we think of good such choices?

5.3 Bayesian Interpretation

Let us suppose θ is a random vector. That will make y in (34) random variables that are correlated with θ . If the prior (before Y has been observed) covariance matrix of θ is P , then it is known that the maximum a posteriori

(after Y has been observed) estimate of θ is given by (36a). [See [2] for all technical details in this section.]

So a natural choice of P is to let it reflect how much is known about the vector θ .

5.4 “Empirical Bayes”

Can we estimate this matrix P in some way? Consider (34). If θ is a Gaussian random vector with zero mean and covariance matrix P , and E is a random Gaussian vector with zero mean and covariance matrix I , and Φ is a known, deterministic matrix, then from (34) also Y will be a Gaussian random vector with zero mean and covariance matrix

$$Z(P) = \Phi P \Phi^T + I \quad (38)$$

(Two times) the negative logarithm of the probability density function (pdf) of the Gaussian random vector Y will thus be

$$W(Y, P) = Y^T Z(P)^{-1} Y + \log \det Z(P) \quad (39)$$

That will also be the negative log likelihood function for estimating P from observations Y , so the ML estimate of P will be

$$\hat{P} = \arg \min W(Y, P) \quad (40)$$

We have thus lifted the problem of estimating θ to a problem where we estimate parameters (in) P that describe the distribution of θ . Such parameters are commonly known as *hyperparameters*.

If the matrix Φ is not deterministic, but depends on E in such a way that row $\varphi^T(t)$ is independent of the element $e(t)$ in E , it is still true that $W(P)$ in (39) will be the negative log likelihood function for estimating P from Y , although then Y is not necessarily Gaussian itself. [See, e.g. Lemma 5.1 in [3].]

5.5 FIR Models

Let us now return to the impulse response (5) and assume it is finite (FIR):

$$G(q, \theta) = \sum_{k=1}^m b_k u(t-k) = \varphi_u^T(t) \theta_b \quad (41)$$

where we have collected the m elements of $u(t-k)$ in $\varphi(t)$ and the m impulse response coefficients b_k in θ_b . That means that the estimation of FIR models is a linear regression problem. All that was said above about linear regressions, regularization and estimation of hyperparameters can thus be applied to estimation of FIR

models. In particular suitable choices of P should reflect what is reasonable to assume about an impulse response: If the system is stable, b should decay exponentially, and if the impulse response is smooth, neighbouring values should have a positive correlation. That means that a typical regularization matrix P^b for θ_b would be matrix whose k, j element is something like

$$P_{k,j}^b(\alpha) = C \min(\lambda^k, \lambda^j); \quad \lambda < 1 \quad \alpha = [C, \lambda] \quad (42)$$

The hyperparameter α can then be tuned by (40):

$$\hat{\alpha} = \arg \min W(Y, P^b(\alpha)) \quad (43)$$

This method of estimating impulse response, possibly followed by a model reduction of the high order FIR model (“`modred(idss(firmodel), n)`”) has been extensively tested in Monte Carlo simulations in [2]. They clearly show that the approach is a viable alternative to the classical ML/PEM methods, and may in some cases provide better models. An important reason for that is that the tricky question of model order determination is avoided.

5.6 ARX Models

Recall that high order ARX models provide increasingly better approximations of general linear systems. We can write the ARX-model (9) as

$$y(t) = -a_1 y(t-1) - \dots - a_n y(t-n) + b_1 u(t-1) + \dots + b_m u(t-m) = \varphi_y^T(t) \theta_a + \varphi_u^T(t) \theta_b = \varphi^T(t) \theta \quad (44)$$

where φ_y and θ_a are made up from y and a in an obvious way. That means that also the ARX model is a linear regression, to which the same ideas of regularization can be applied. Eq (44) shows that the predictor consists of two impulse responses, one from y and one from u and similar ideas on the parameterization of the regularization matrix can be used. It is natural to partition the P -matrix in (36a) along with θ_a, θ_b and use

$$P(\alpha_1, \alpha_2) = \begin{bmatrix} P^a(\alpha_1) & 0 \\ 0 & P^b(\alpha_2) \end{bmatrix} \quad (45)$$

with $P^{a,b}(\alpha)$ as in (42).

5.7 Related work

The text in this section essentially follows [2]. Important contributions of the same kind, based on ideas from machine learning, have been described in [11] and [10].

6 Nonparametric Models of Linear Systems

Classical non-parametric models of linear system are methods to estimate the frequency functions $G(e^{i\omega}), H(e^{i\omega})$ in (6) directly from data (often the ETFE (19)) without first finding any model parameters.

$$\Phi_y(\omega) = \lambda |H(e^{i\omega})|^2 \quad (46)$$

is the *spectrum* of the disturbances, such methods are often referred to as *spectral analysis*.

6.1 Classical Spectral Analysis

Classical spectral analysis is a way of directly smoothing the ETFE by averaging over a window sliding across it:

$$\hat{G}(e^{i\omega}) = \frac{\int_{-\pi}^{\pi} W_{\gamma}(\xi - \omega) \beta(\xi) \hat{G}_N(e^{i\xi}) d\xi}{\int_{-\pi}^{\pi} W_{\gamma}(\xi - \omega) \beta(\xi) d\xi} \quad (47)$$

here β is a weighting that may account for the varying reliability of \hat{G} over the frequencies. W_{γ} is the window which performs the smoothing. γ is a parameter that governs the width of the window which decides the trade-off between frequency resolution and noise-sensitivity. This is a variant of the fundamental bias-variance trade off which is present in all estimation problems. See Section 6.4 in [3] for more details around this.

6.2 Local Polynomial Techniques

Quite recently, an alternative way of smoothing the ETFE has been suggested, [13]. Consider the frequency measurements (16). Assume that they have been collected on an equidistant grid, and denote for simplicity

$$Y(k) = Y_N(e^{i\omega_k}); \quad U(k) = U_N(e^{i\omega_k}); \quad G_k = G(e^{i\omega_k}) \quad (48)$$

They are related to the frequency function as

$$Y(k) = G_k U(k) + T_k + E_k \quad (49)$$

where T_k are transient errors and E_k is noise. If G_k and T_k were constant over a certain frequency interval, they could easily be estimated by averaging over the more rapidly changing noise term. Now assume that the frequency function and the transient error change rather slowly with k , so that they are not constant but may

vary with frequency like a low order (p) polynomial:

$$\begin{aligned} G_{k+r} &= \sum_{j=0}^p x_j r^j \\ T_{k+r} &= \sum_{j=0}^p y_j r^j \\ \beta_k &= [x_j, y_j, j = 0, \dots, p] \end{aligned} \quad (50)$$

Then we could use the model (49) over a frequency range around k of more observations ($2M + 1$) than the number of unknown parameters ($2p + 2$) and estimate the parameters by least squares:

$$\hat{\beta}_k = \arg \min_{\beta_k} \sum_{r=-M}^M |Y(k+r) - G_{k+r} U(k+r) - T_{k+r}|^2 \quad (51)$$

The central estimate, $r = 0$, will then be our estimate:

$$\hat{G}(e^{i\omega_k}) = G_{k+0} = \hat{x}_0 \quad (52)$$

This Least Squares estimation has to be performed at each frequency value ω_k of interest. The calculations are thus more extensive than for the classical estimate (47), but it is shown in [13] that the accuracy is much better. An interesting alternative is to use rational rather than polynomial approximations in (50) as discussed and illustrated in [9].

7 Conclusions

System Identification is an area of clear importance for practical systems work. It has now a well developed theory and is a standard tool in industrial applications. Even though the area is quite mature with many links to classical theory, new exciting and fruitful ideas keep being developed. This article has tried to illustrate both these aspects. Further discussions and views on the current status and future perspectives on system identification are given in e.g. [4] and [7].

References

- [1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19:716–723, 1974.
- [2] Tianshi Chen, Henrik Ohlsson, and Lennart Ljung. On the estimation of transfer functions, regularizations and Gaussian processes-Revisited. *Automatica*, 48(8):1525–1535, 2012.
- [3] L. Ljung. *System Identification - Theory for the User*. Prentice-Hall, Upper Saddle River, N.J., 2nd edition, 1999.
- [4] L. Ljung. Perspectives on system identification. *IFAC Annual Reviews*, Spring Issue, 2010.

- [5] L. Ljung and B. Wahlberg. Asymptotic properties of the least-squares method for estimating transfer functions and disturbance spectra. *Adv. Appl. Prob.*, 24:412–440, 1992.
- [6] Lennart Ljung. Prediction error estimation methods. *Circuits, systems, and signal processing*, 21(1):11–21, 2002.
- [7] Lennart Ljung, Hakan Hjalmarsson, and Henrik Ohlsson. Four Encounters with System Identification. *European Journal of Control*, 17(5-6):449–471, 2011.
- [8] Lennart Ljung and Adrian Wills. Issues in sampling and estimating continuous-time models with stochastic disturbances. *AUTOMATICA*, 46(5):925–931, 2010.
- [9] T. McKelvey and G. Guerin. Nonparametric frequency response estimation using a local rational model. In *Proc. 18th IFAC Symposium on System Identification*, Brussels, Belgium, July 2012.
- [10] G. Pillonetto, A. Chiuso, and G. De Nicolao. Prediction error identification of linear systems: a nonparametric Gaussian regression approach. *Automatica*, 47(2):291–305, 2011.
- [11] G. Pillonetto and G. De Nicolao. A new kernel-based approach for linear system identification. *Automatica*, 46(1):81–93, January 2010.
- [12] R. Pintelon and J. Schoukens. *System Identification – A Frequency Domain Approach*. IEEE Press, New York, 2nd edition, 2012.
- [13] J. Schoukens, G. Vandersteen, K. Barbe, and R. Pintelon. Nonparametric preprocessing in system identification – a powerful tool. *European Journal of Control*, 15:260–274, 2009.
- [14] Y.C. Zhu. Asymptotic properties of prediction error methods. *Int. J. Adaptive Control and Signal Processing*, 3:357–373, 1989.
- [15] Y.C. Zhu and A. C. M. P. Backx. *Identification of Multivariable Industrial Processes for Diagnosis and Control*. Springer Verlag, London, 1993.