Overview of System Identification

Dr.-Ing. Sudchai Boonto
Assistant Professor

August 24, 2017

Department of Control System and Instrument Engineering, KMUTT

Identifiability

General Concepts

The ability to identify a *unique* model for a given system depends on three critical aspects:

- **Model:** Whether there exists a unique mapping between the model and the parameters being estimated — **Model Identifiability**
- **Experimental conditions:** Whether the input has generated the requisite information required to distinguish between two candidate models — **Input Identifiability** (Informative enough, rich enough)
- **Estimation method:** Whether the estimation method is capable of estimating the "true" parameters if infinite samples are available, which is termed as an **asymptotic property** of the estimator. The technical term is **consistency**.

We will discuss the theoretical definition of the identifiability later.

Uniqueness

Consider fitting the model $y[k, \theta] = \theta_1 \theta_2 u[k]$ to a given data, where $u[k]$ and $y[k]$ are the input and output of a system, while $\hat{\theta} = [\theta_1 \ \theta_2]^T$ is the parameter vector to be identified.

The prediction of this model to a given input is

$$\hat{y}[k, \hat{\theta}] = \theta_1 \theta_2 u[k]$$

The two different parameter values of $\hat{\theta}_1$ and $\hat{\theta}_2$ produce identical predictions. Then, we have

$$\hat{y}[k, \hat{\theta}_1] = \hat{y}[k, \hat{\theta}_2] \implies \hat{\theta}_1 = \hat{\theta}_2$$

The model (predictor) space is not one-to-one. Formally, the model is said to be not (globally) identifiable. Consequently, it is not possible to arrive at a unique estimate of $\hat{\theta}$.

Uniqueness

On the other hand, if the model is re-parametrized in terms of a single parameter $\beta = \theta_1\theta_2$, then the model is identifiable at all points in the β space.

The above example suggests that **re-parametrization of a model**, in this case from a higher dimensional to a lower-dimensional parameter space, can improve identifiability for that model.

good information

Consider a linear time-invariant (LTI) system governed by the following input-output relationship (3^{rd}-order finite impulse response system):

$$y[k] = b_1 u[k-1] + b_2 u[k-2] + b_3 u[k-3],$$

with $b_1 = 1$, $b_2 = 0.6$, and $b_3 = 0.3$. Suppose a sinusoidal input of the form $u[k] = \sin(2\pi(0.1)k) = \sin(\omega_0 k)$ is applied to the system. Under this input, we have

$$\begin{aligned} y[k] &= b_1 \sin(\omega_0(k-1)) + b_2 \sin(\omega_0(k-2)) + b_3 \sin(\omega_0(k-3)) \\ &= \left(b_1 + \frac{b_2}{2\cos\omega_0}\right) \sin(\omega_0 k - \omega_0) + \left(b_3 + \frac{b_2}{2\cos\omega_0}\right) \sin(\omega_0 k - 3\omega_0) \\ &= \tilde{b}_1 \sin(\omega_0 k - \omega_0) + \tilde{b}_3 \sin(\omega_0 k - 3\omega_0) \end{aligned}$$

good information

Thus, a 3-parameter model manifests as a 2-parameter model when viewed through the lens of a mono-frequency input. Unfortunately, it is not possible to uniquely recover b_1, b_2, and b_3 from $\tilde{b}_1$ and $\tilde{b}_3$.

- the above example is formally termed as loss of identifiability due to insufficient information, which here is due to lack of sufficient input excitation.
- The input that contains enough frequency components is $u[k] = \sin(\omega_0 k) + \sin(\omega_1 k)$.

Note:

$$b_2 \sin(\omega_0 k - 2\omega_0 + \omega_0) = b_2 \sin(\omega_0 k - 2\omega_0)\cos\omega_0 + b_2 \sin\omega_0 \sin(\omega_0 k - 2\omega_0)$$

$$b_2 \sin(\omega_0 k - 2\omega_0 - \omega_0) = b_2 \sin(\omega_0 k - 2\omega_0)\cos\omega_0 - b_2 \sin\omega_0 \sin(\omega_0 k - 2\omega_0)$$

$$b_2 \sin(\omega_0 k - \omega_0) + b_2 \sin(\omega_0 k - 3\omega_0) = 2b_2 \sin(\omega_0 k - 2\omega_0)\cos\omega_0$$

Signal-to-Noise Ratio

Signal-to-noise ratio

The stochastic effects in the measurements may be too high to detrimental to model quality.

- Choosing sampling rates much faster than the pace at which outputs change can bring in more noise than actual process variation.

The **Signal-to-noise ratio (SNR)** is defined by

$$\text{SNR} = \frac{\text{Variance of signal}}{\text{Variance of noise}}$$

- The term **signal** refers to the **true response** of the system.
- Having a high SNR is critical to obtaining reliable parameter estimates, regardless of the estimation method.
- The lower the SNR, the more ambiguous is the estimate of the input-output model.

Effect of SNR on Parameter Estimation

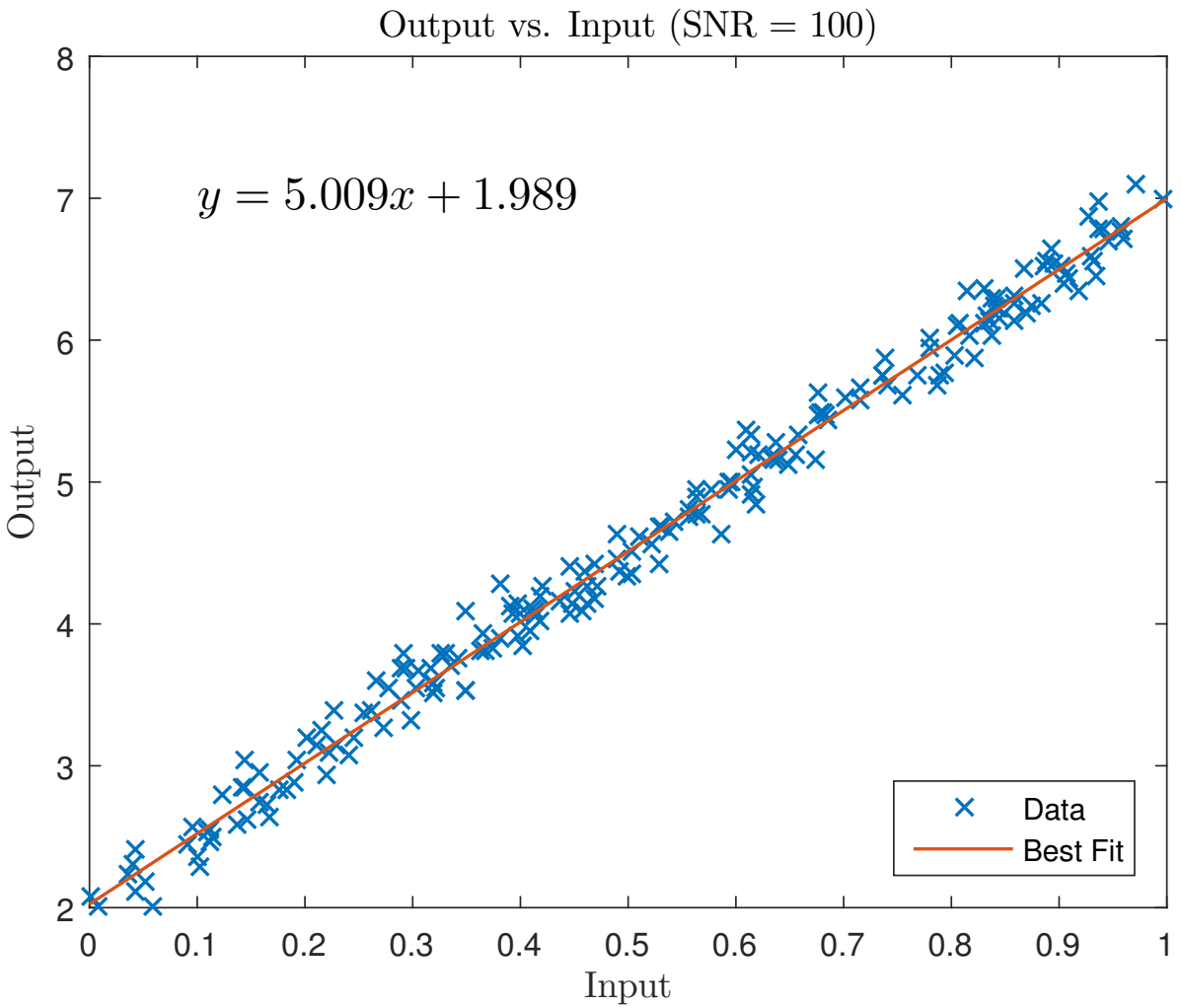
Assuming the relationship between the output $y[k]$ and the input $u[k]$ of a system is known to be

$$y[k] = b_1 u[k] + b_0,$$

where $b_1 = 5$ and $b_0 = 2$. The input and the measured output $y_m[k] = y[k] + v[k]$ is available, where the measurement error $v[k]$ is assumed to be random. The best linear fit and the parameter estimates fro two different settings of SNR $= \sigma_y^2/\sigma_v^2$, namely (i) SNR $= 100$ and (ii) SNR $= 10$ obtained from $N = 200$ samples of $(y_m[k],\ u[k])$ data.

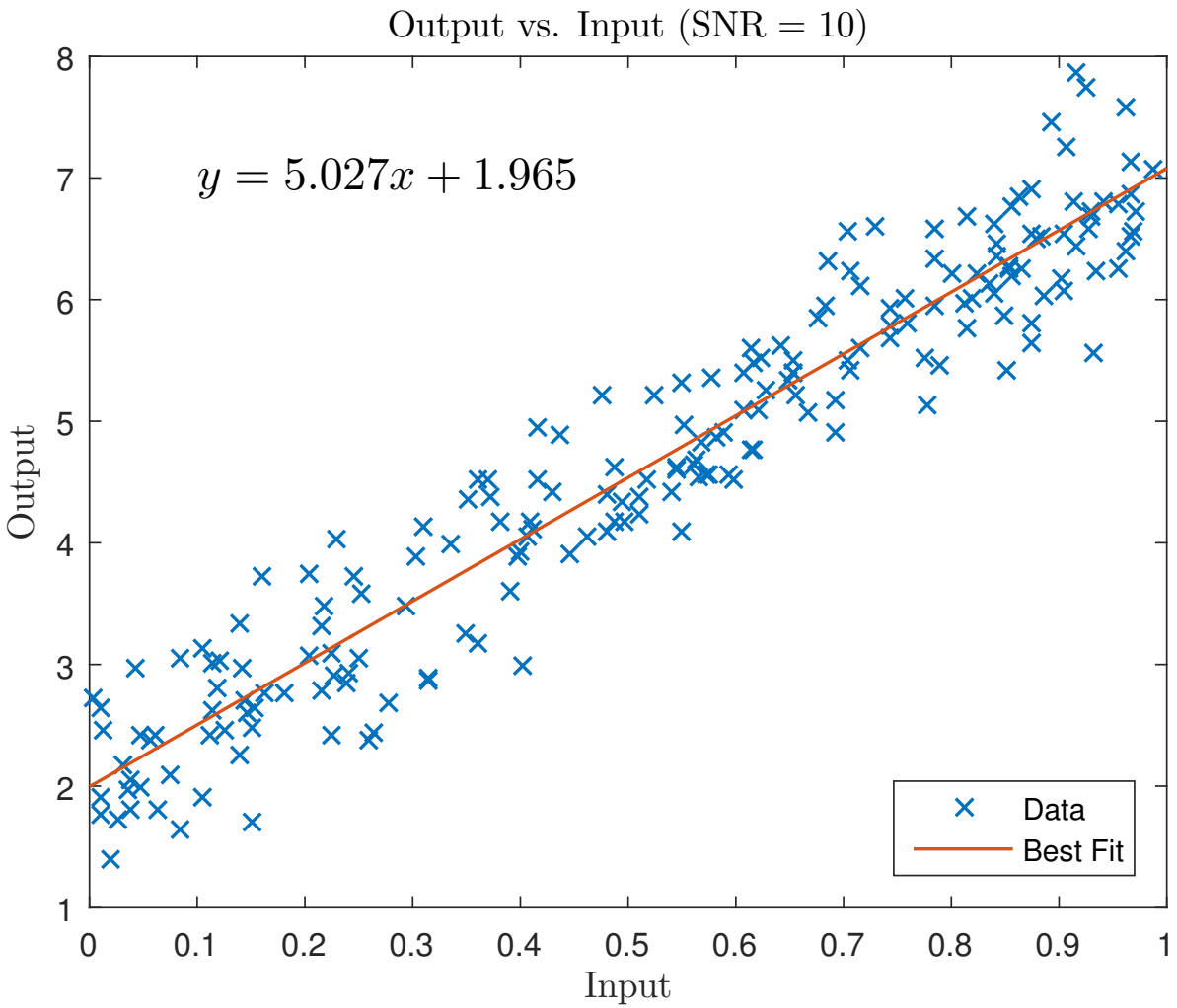
- The estimates do not vary significantly with the change in SNR, the errors in $\hat{b}_i$. The $\sigma_{\hat{b}_i}$ in the estimates increase roughly by a factor of three.
- The increase is theoretically given by $\sqrt{100/10} = 3.162$. The lower the SNR, the lower the reliability (confidence) of the resulting parameter estimate.

Effect of SNR on Parameter Estimation



$$\sigma_{\hat{b}_1} = 0.036, \qquad \sigma_{\hat{b}_2} = 0.02$$

Effect of SNR on Parameter Estimation



$$\sigma_{\hat{b}_1} = 0.114, \qquad \sigma_{\hat{b}_2} = 0.064$$

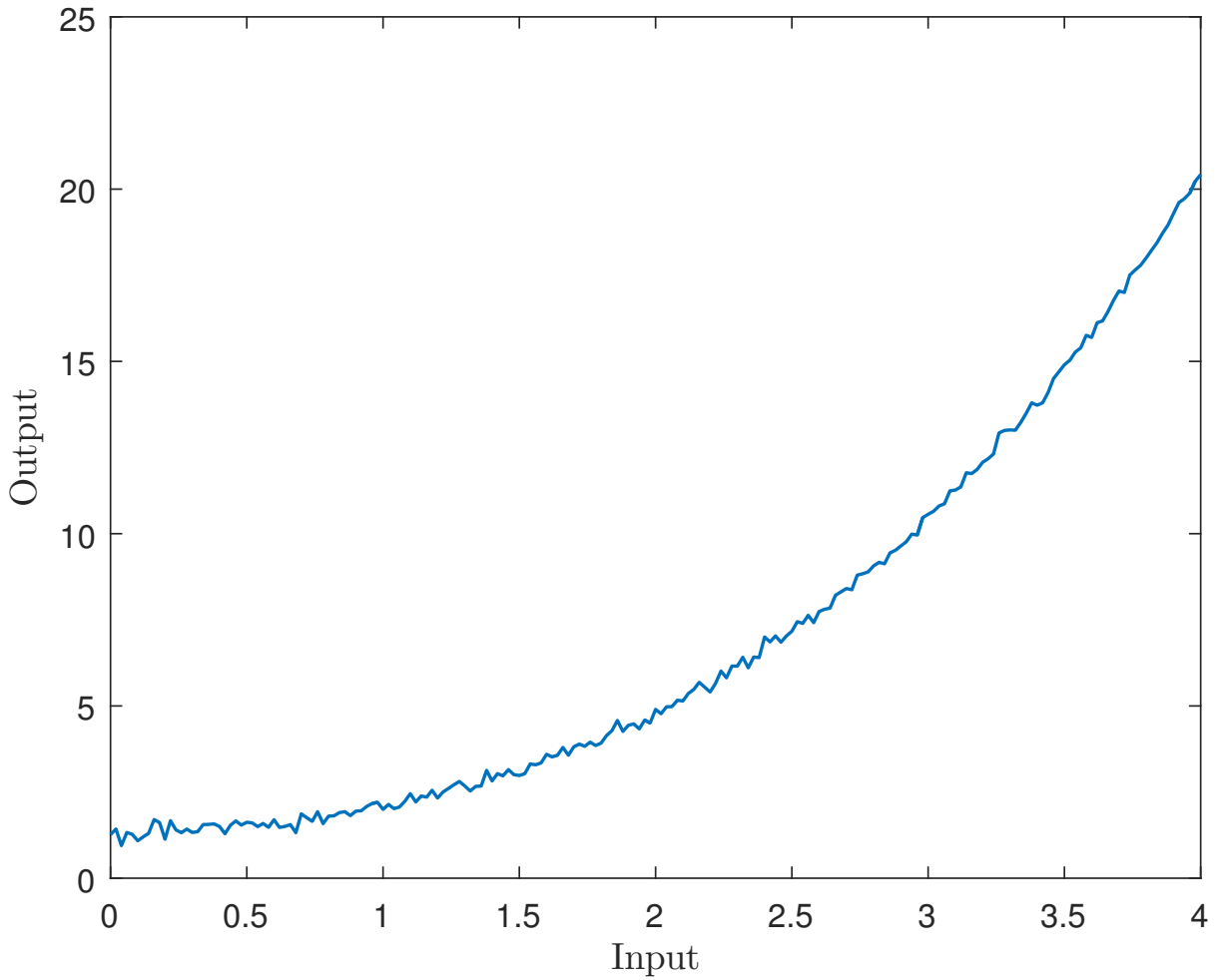
Overfitting

Overfitting

Overfitting occurs when the model is trained to capture the **local** feature of the data rather than the **global** characteristics.

- the situation arises when one misconstrues the stochastic effects in the data as a part of the deterministic (input-output) effects, i.e., when the chance variations in the response are attributed to the changes in the input variables.
- This situation occurs when the user **over-specifies** the complexity of the deterministic portion in a bid to explain the output as accurately as possible.
- the benefit from increasing the complexity of the deterministic model is the improved fit on the training data, i.e., lower prediction error.
- the reduction in the bias (of the prediction) comes with the risk of high standard errors (variance) in the model (parameter) estimates.

Overfitting



It is clear that a polynomial fit can capture the relationship reasonably well.

Overfitting

The true model used for data generation is

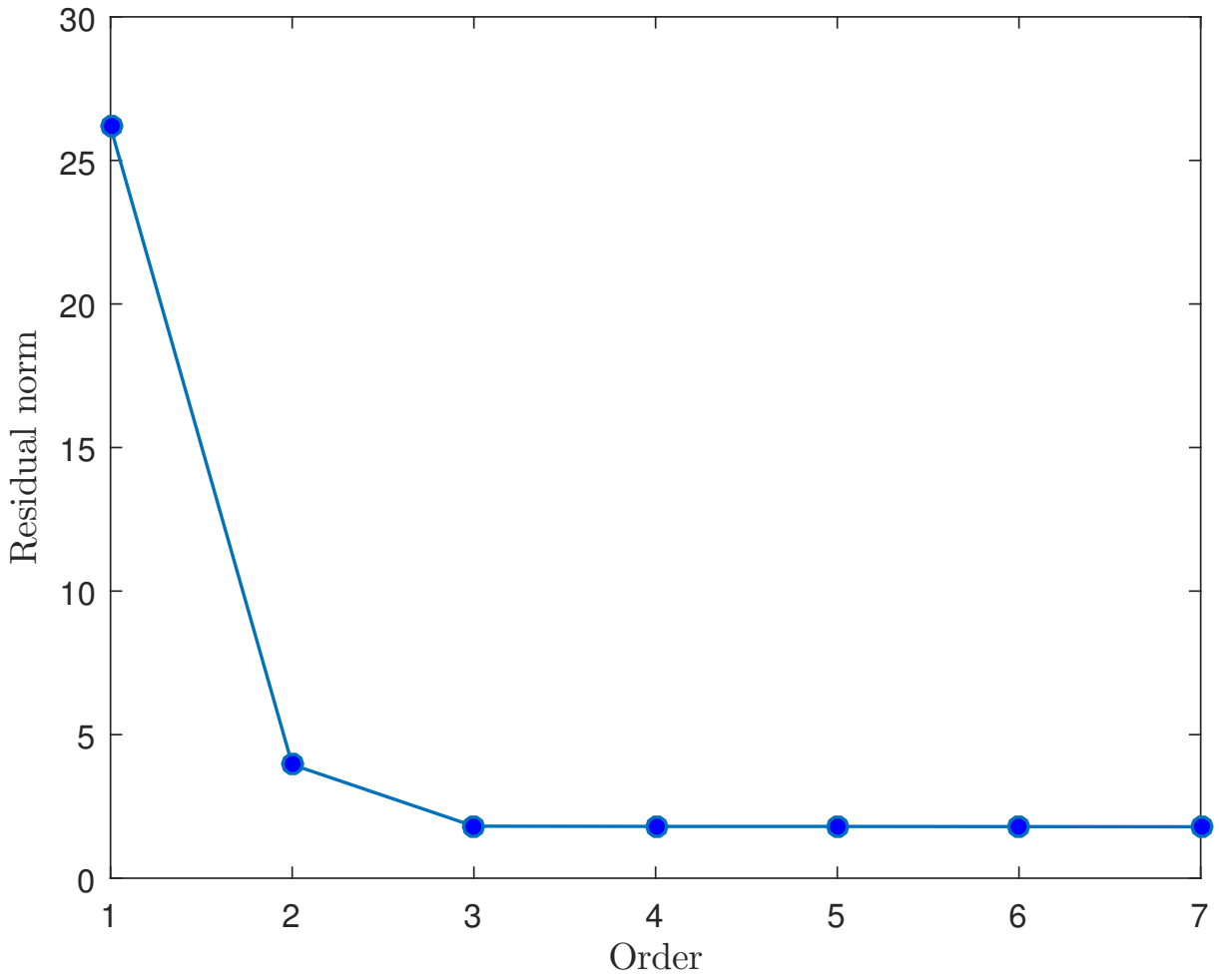
$$y[k] = 1.2 + 0.4u[k] + 0.3u^2[k] + 0.2u^3u[k] + v[k],$$

where $v[k]$ is an ideal random noise (unpredictable stochastic signal) such that the SNR is set to 10.
The predictor (fitted polynomial) is

$$\hat{y}[k] = a_0 + a_1u[k] + a_2u^2[k] + a_3u^3[k] + \ldots + a_nu^n[k].$$

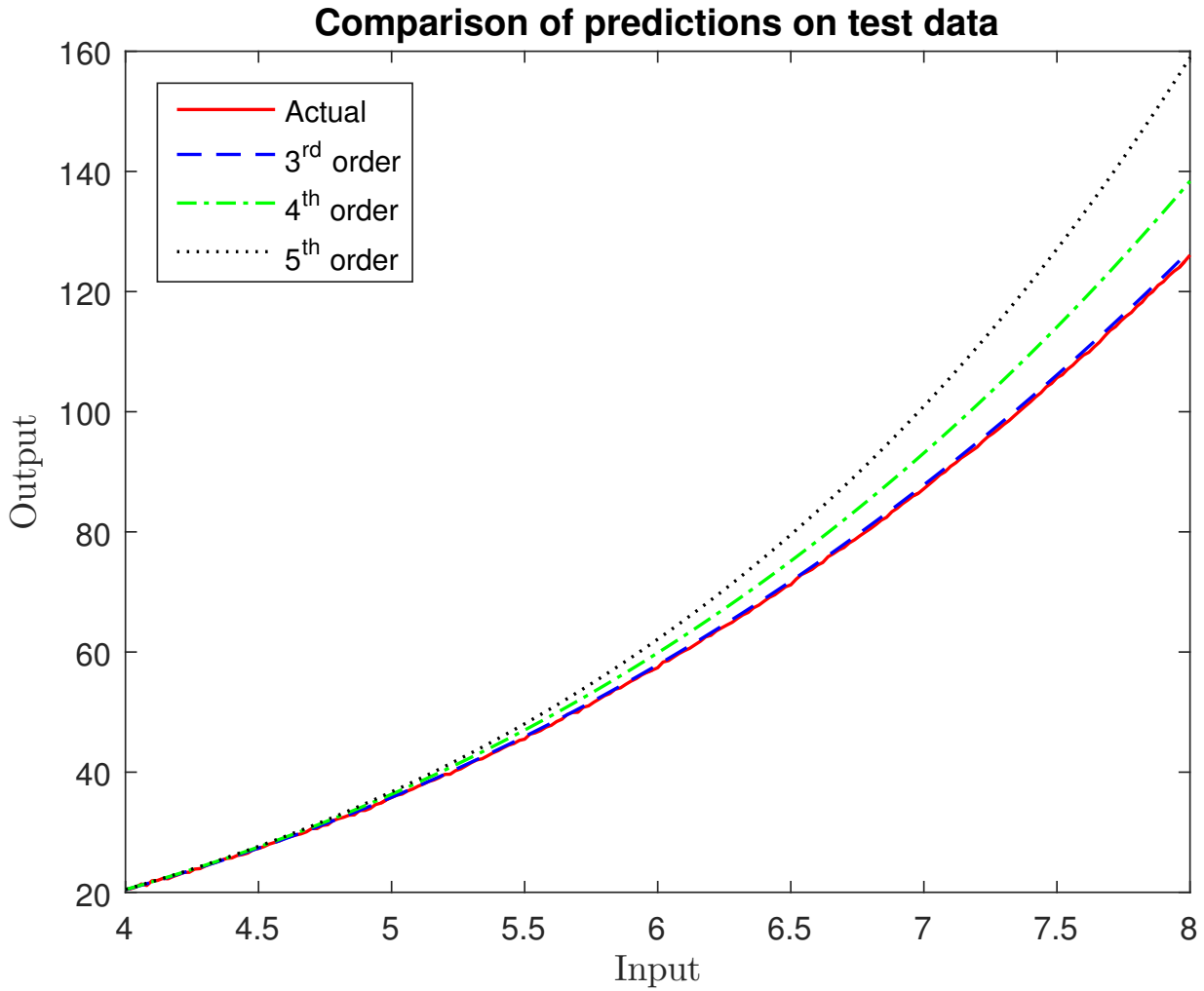
From the orthogonality, we know that the more terms the more accurate.

Overfitting



Here, the sum squared error of residual is drafted against the order of the polynomial.

Overfitting



The best approximation:

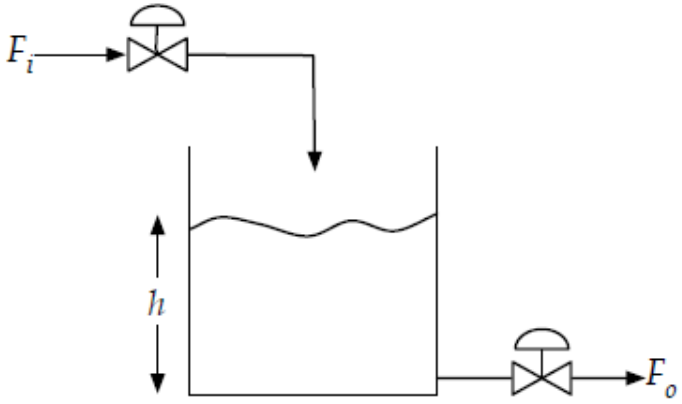
$$\hat{y}[k] = 1.183 + 0.384u[k] + 0.314u^2[k] + 0.198u^3[k]$$

Overfitting

- The fourth and fifth order models negligibly lower residual norms on the training data and therefore fail miserably on the test data.
- In fact the fifth-order model produces unstable predictions.
- The instability of predictions is one of the perils in overfitting, which can be avoided by examining the errors in parameter estimates of these models in conjunction with plot.
- There exist no strict rules that can completely prevent the overfitting phenomenon.

Example

Example: Liquid Level System



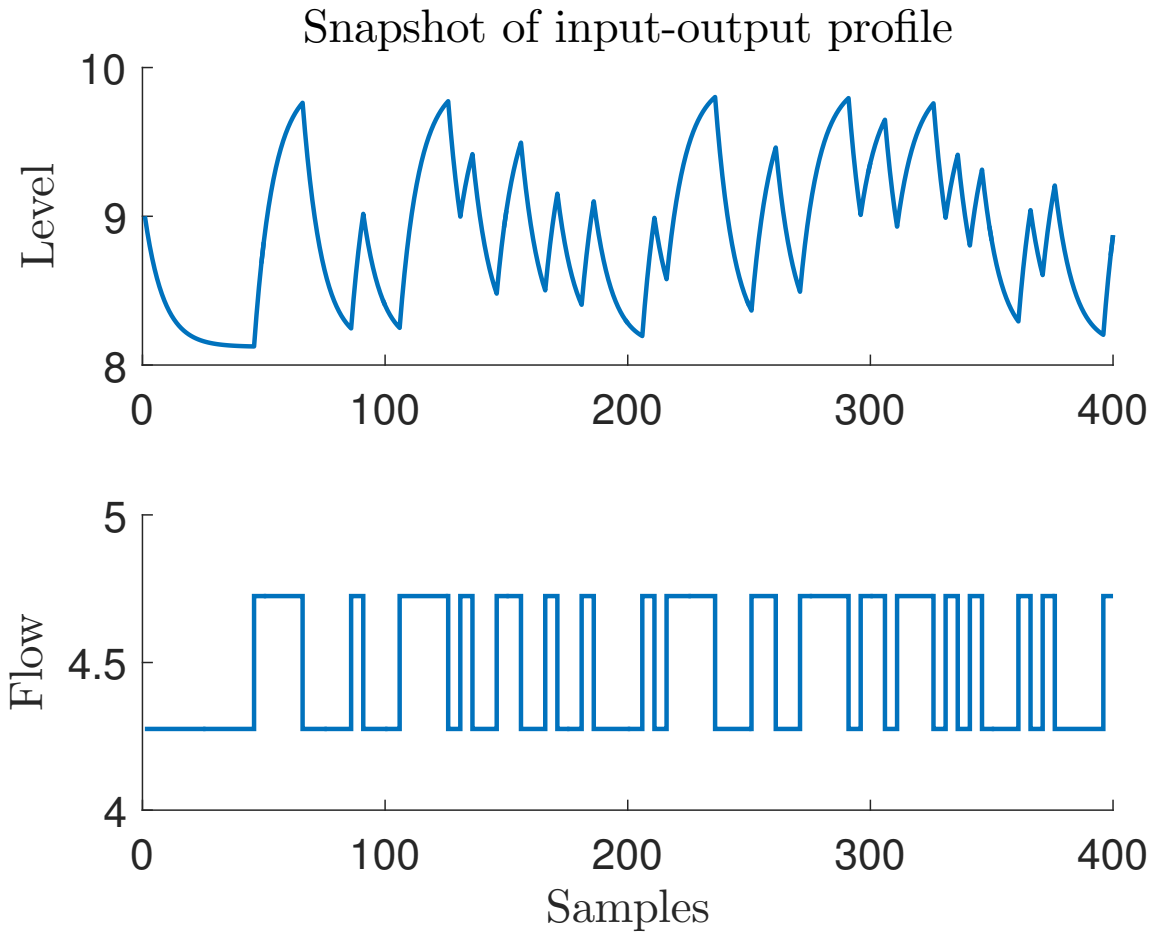
The model for simulation is

$$\frac{dh(t)}{dt} + \frac{1}{A_c} C_v \sqrt{h(t)} = \frac{1}{A_c} F_i(t)$$

$$F_o(t) = C_v \sqrt{h(t)}$$

Here C_v is the valve coefficient at the output. The quantity A_c is the cross-sectional area of the cylindrical tank. The system is brought to a steady state before exciting it with the designed input. With the operating conditions set to $F_i(t) = 4.5$ cu. ft. /min., $C_v = 1.5$ and $A_c = 0.5$ ft^2, the nominal level is 9 ft.

Example: Liquid Level System



The input is PRBS with sampling of $T_s = 1$ min. The output is added by noise for realistic simulation such that the output SNR is st to a value of 10.

Data Visualization and Preliminary Analysis

A First step:

- see any drifts, outliers, etc.
- From the previous plot, there is no any polynomial trends and other anomalies of concern (such as outliers, missing data).
- Steady-state can be determined experimentally before introducing changes to the input. When such experiments have not been performed , an alternative is to use the average of the readings as a nominal operating point.
- A simple mean centering operation of the input-output data is used to generate the required deviation variables.

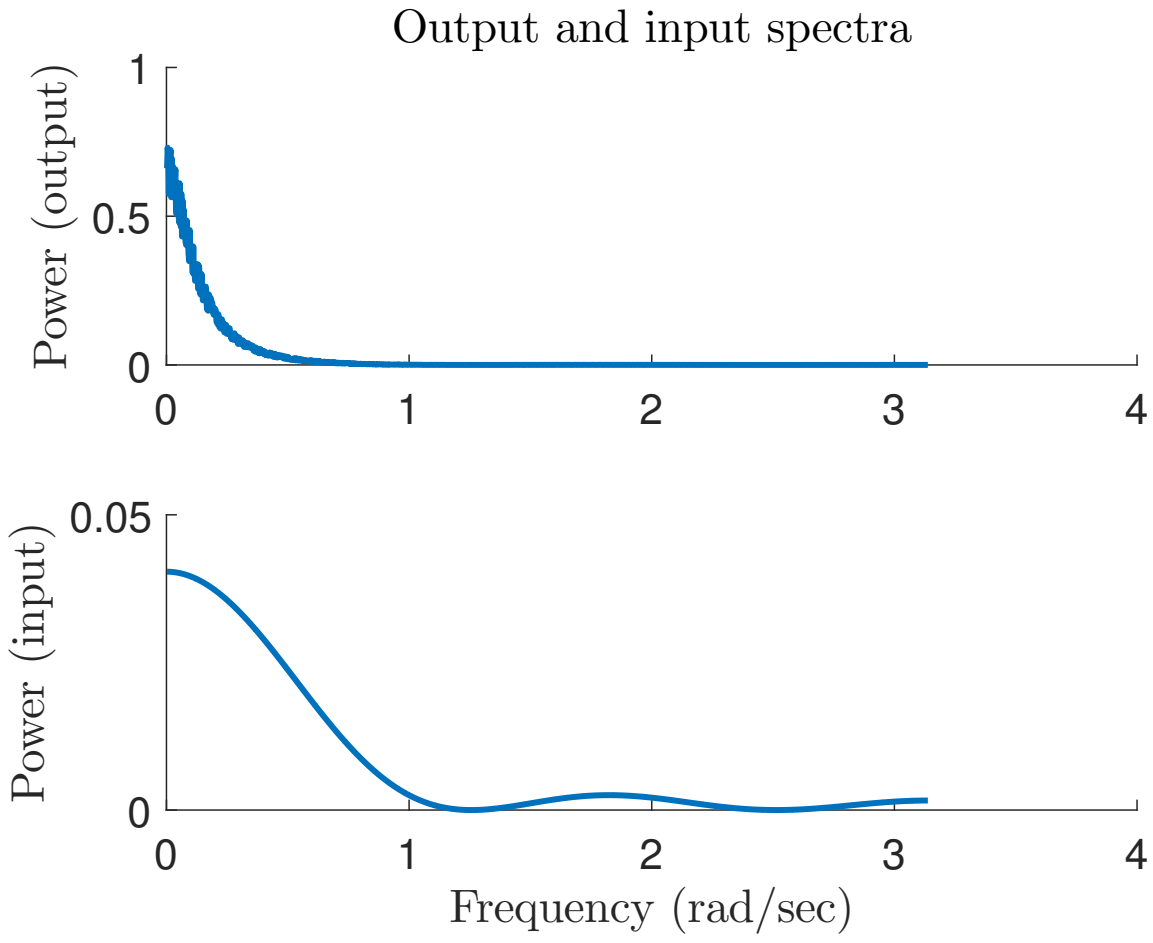
$$y[k] = \tilde{y}[k] - \bar{y}; \qquad u[k] = \tilde{u}[k] - \bar{u},$$

where variables with ˜ are absolute-valued and the quantities under the bar are average of the respective variables.

Spectral analysis

- It is useful to examine the frequency content of the output signal so as to obtain insights into the **filtering** nature of the process.
- The **spectral plot** consisting of a plot of power vs. frequency is used for this purpose.
- High power in a frequency band implies the strong presence of those frequency component

Spectral Analysis



Most of the input power has been packed in the low- to mid-frequency band. This is in fac a part of the input design strategy because liquid level systems are low-pass filters.

Partitioning the data

For the modelling purpose, the data is partitioned into two sets:

- one partition consisting of the first $N = 1500$ samples for training the model.
- the second set consisting of the remainder of the data, used in cross-validation of the model.
- Both data sets are expressed in terms of deviation variables with the nominal operating point determined from the training data.

Building Non-Parametric Models

The non-parametric models provide insights into several important (deterministic) process characteristics with minimal assumptions:

- **Time-delay:** The impulse response (cross-correlation) method is the classical approach for delay estimation.
- **Gain:** Step response coefficients are ideally suited for estimating gain.
- **Time-constant:** Step response is naturally suited for estimating the time-constant parameter.

Impulse Response estimates

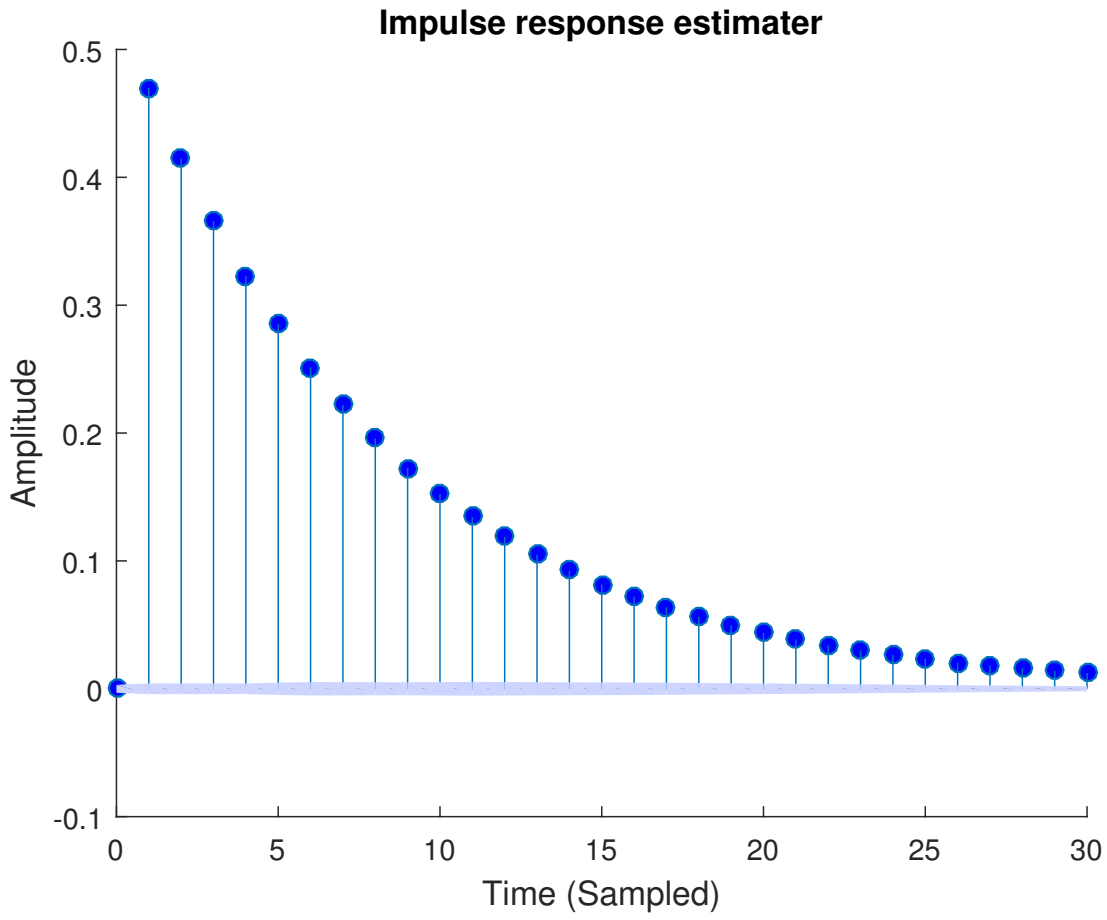
The impulse response (IR) estimates are obtained by fitting a finite-length impulse response (FIR) model:

$$y[k] \approx \sum_{l=0}^{M} g[l]u[k-l]$$

to the data using the **least squares** method.

- Whe the system has a delay of D samples, then the first D of IR coefficients are identically zero.
- The corresponding estimated coefficients will be however small non-zero values.
- In practice, a test of significance is performed wherein the estimated coefficients lower than a statistically determined threshold are termed as insignificant.

Spectral Analysis

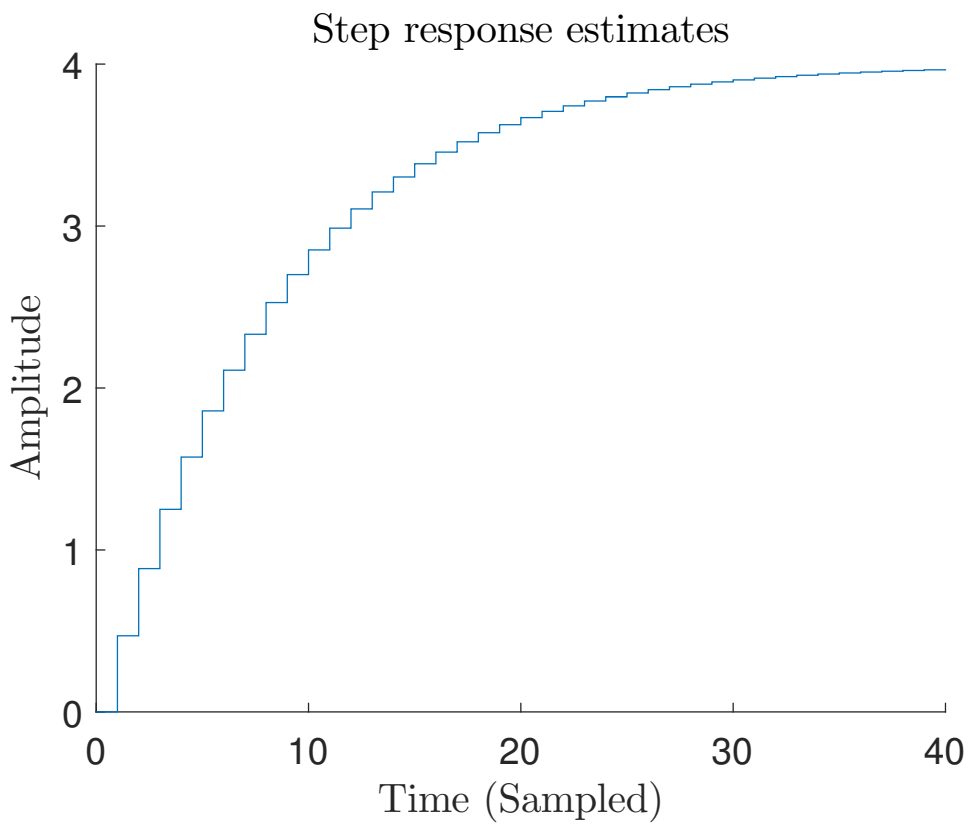


The decaying nature of the IR estimates is a clear indication that the sampled-data system is **stable**, which agrees very well with our physical knowledge of the process.

Step response model

Estimates of unit step response coefficients are obtained from the IR coefficients using a simple relationship

$$y_s[k] = \sum_{n=1}^{k} g[n]$$



Step response model

The resulting estimates:

- a first-order (or an overdamped higher-order) dynamics with a gain of approximately 3.7 units.
- If the system is approximated as a first-order process, the time-constant is about 7 samples (minutes)
- the system is stable.

Building Parametric Models

Building Parametric Models

The objective is to identify a difference equation (DE) model for the **deterministic** process from the given data. The user need to give a suitable value of the dalay-time and order of the DE form.

The first-order difference equation with a delay of 1 unit sample for the **deterministic** process is

$$x[k] + a_1 x[k-1] = b_1 u[k-1],$$

where $x[k]$ denotes the **unobserved** true discrete-time (deterministic) response of the process.
The parameters $\theta = [a_1 \ b_1]^T$ are estimated such that the sum squared one-step ahead prediction errors is minimized,

$$\min_{\theta} \sum_{k=0}^{N-1} \left(x[k] - \hat{x}[k|k-1]\right)^2$$

where

$$\hat{x}[k|k-1] = -a_1 x[k-1] + b_1 u[k-1]$$

is the **prediction** of $x[k]$ given the knowledge of $x[\cdot]$ and $u[\cdot]$ until the $(k-1)^{\text{th}}$ instant.
However, this is only possible when the true response is known. In reality, only a measurement of $x[k]$ is available. Therefore, it is natural re-write the minimization in terms of the **measurement prediction error:**

$$\min_{\theta} \sum_{k=0}^{N-1} \left(y[k] - \hat{y}[k|k-1]\right)^2,$$

where $\hat{y}[k|k-1]$ is the prediction of the measurement $y[k]$ given the knowledge of measurements and inputs until the $(k-1)^{\text{th}}$ instant. Later, we will use $\hat{y}[k]$ to denote $\hat{y}[k|k-1]$.

Building Parametric Models

To construct the overall model for $y[k]$, we first assume that stochastic effects, collectively denoted by $v[k]$, are additive

$$y[k] = x[k] + v[k]$$

Naturally, the (one-step ahead) prediction of the measurement is

$$\hat{y}[k] = \hat{x}[k] + \hat{v}[k],$$

where $\hat{v}[k]$ is the (one-step ahead) prediction of the disturbances and noise.

- the deterministic signal $x[k]$ is modeled by the difference. A model for $v[k]$ is required to complete the picture.
- Different models exist depending on the assumptions made on the predictability of $v[k]$, leading to different descriptions for $y[k]$.

Output-error model

A simple assumption is that the **error in the measurement is absolutely unpredictable** given any amount of past. For zero-mean noise this mean

$$\hat{v}[k] = 0$$

From the first-order model, we have

$$\hat{y}[k] = \hat{x}[k] = -a_1 x[k-1] + b_1 u[k-1].$$

The unknown variables are a_1, b_1, and $x[k-1]$ (never know because of noise). It is clear that the predictor is **non-linear in unknowns** ($a_1 x[k-1]$). Since the white-noise error directly enters the output, the model producing the predictor above is termed as the **output-error** model.

equation-error model

The assumption is that $v[k]$ is **predictable**, but with the additional requirement of a **linear predictor** for the measurement. From the idea, we have

$$y[k] = -a_1 y[k-1] + b_1 u[k-1] + \underbrace{v[k] + a_1 v[k-1]}_{w[k]}$$

$$\hat{y}[k|k-1] = -a_1 y[k-1] + b_1 u[k-1] + \hat{w}[k|k-1]$$

In order to have a **linear predictor** for the measurement $y[k-1]$, we require $\hat{w}[k|k-1] = 0$ meaning $w[k]$ is the white-noise signal $e[k]$. It follows that

$$\hat{y}[k] = -a_1 y[k-1] + b_1 u[k-1]$$
$$\hat{v}[k] = -a_1 v[k-1]$$

leading to a first-order **auto-regressive** predictor for $v[k]$.

equation-error model

Since the error in the difference equation for the measurement is white, we call this description as **equation-error** model, specifically the auto-regressive eXogenous (ARX) model.

Notice that the model for the measurement error has a **coefficient identical to the one for the deterministic response**. A major implication of this method is that the deterministic and noise process share the same dynamics.

Goodness of the model

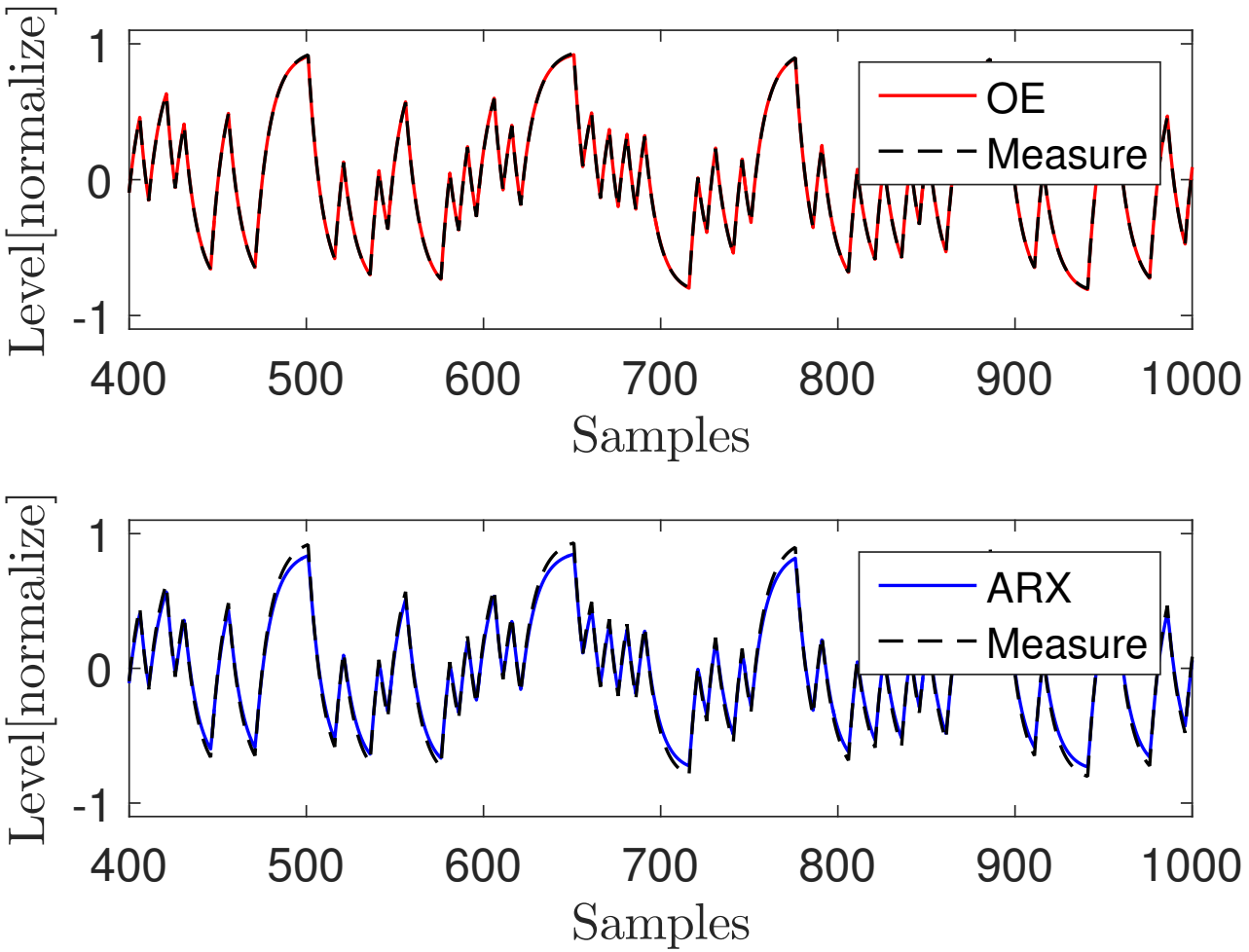
The prediction error, technically termed as **residual**, serves as the key quantity of interest in assessing the goodness of the model. It is formally defined as

$$\varepsilon[k] = y[k] - \hat{y}[k]$$

A good model should not leave behind residuals (from training) that offer further scope for predictions, while avoiding overfitting. Consequently, the following have to be fulfilled:

- the residuals cannot be explained (predicted) by the input (test for the deterministic model).
- the residuals cannot be predicted using its own past, i.e., it is truly unpredictable (test for the stochastic model), and
- the errors in parameter estimates are small or negligible relative to the estimates themselves (test for over-parametrization).

Prediction Analysis

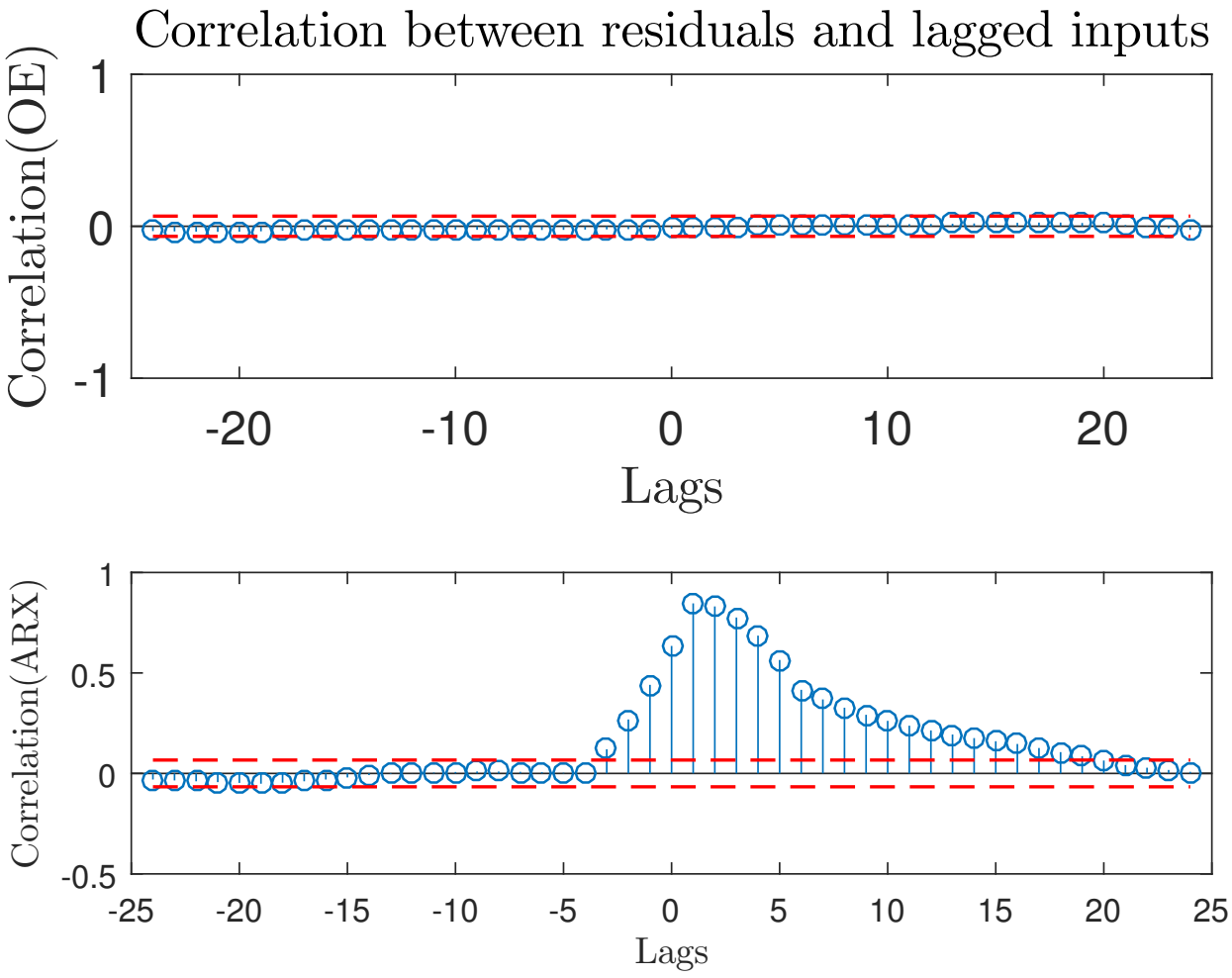


Here, the output-error model fares better than its equation-error counterpart in this respect.

Correlating residuals with inputs

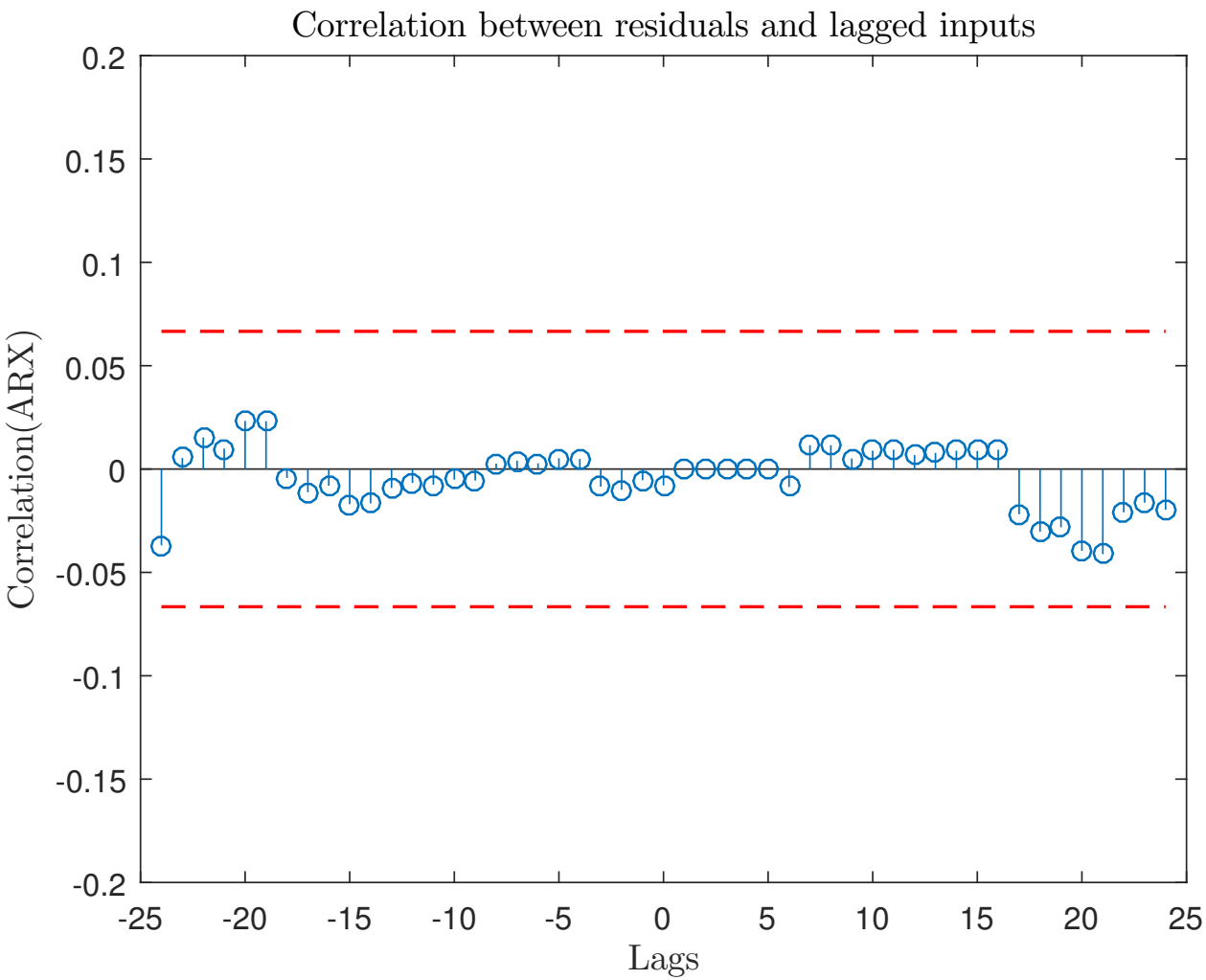
- The foregoing prediction analysis puts the models in close contest.
- To test whether these model leave behind any unexplained input effects, the correlation between the residuals and the **lagged** (time-shifted) inputs are computed for each of these models. This test is also known as **cross-correlation function**.
- A significan correlation between residuals $\varepsilon[k]$ and input $u[k]$ at positive lags directly implies that the effects of input on the process response have not been completely explained.

Correlating residuals with inputs



It is clear that the residuals from the first-order ARX model are significantly correlated with inputs implying that it has not managed to adequately capture the deterministic effects.

Correlating residuals with inputs



For ARX, we can improve it by using fifth-order model. In this case, the OE model is the winner in two respects, namely, parameter estimation error and parsimony.

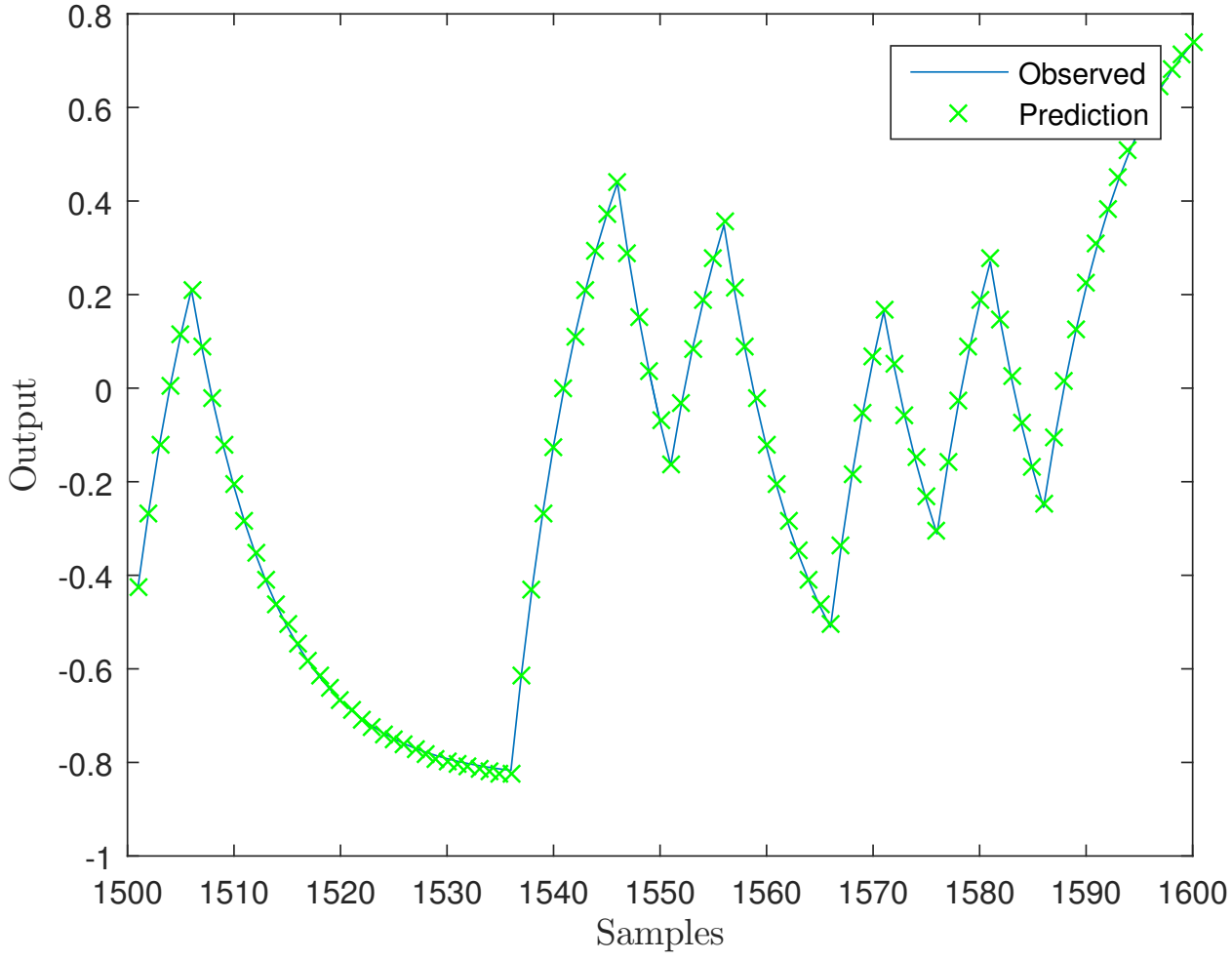
Test for noise model

The next step is to access the stochastic part of the model has satisfactorily explained the random effects. This can be done by plotting the **auto-correlation function** of the residuals.

- It is essentially the correlation between any two samples separated in time by a lag l.
- Any predictability in the sequence manifests as non-zero correlation at a noon-zero lab l.

Cross-Validation

A good model is one which yields good predictions on a fresh data set.



Fitness Test

A common measure of goodness-of-predictions is the normalized root mean square (NRMS) measure of fit:

$$\%R_f = 1 - \frac{\|y - \hat{y}\|_2}{\|y - \bar{y}\|_2} \times 100$$

In this example we have

$$R_f = 98.98\%$$

Final Model

Based on the results of the model assessment tests, namely, the residual analysis, analysis of errors in estimates and cross-validation, the first-order OE model with a delay of unit sample is deemed as the most appropriate:

$$
\begin{aligned}
y[k] &= x[k] + e[k] \\
x[k] &= -\hat{a}_1 x[k-1] + \hat{b}_1 u[k-1] \\
\hat{a}_1 &= -0.8826(\pm 0.0019),\ \hat{b}_1 = 0.4621(\pm 0.0052)
\end{aligned}
$$

Another from of the model is a **transfer function (TF) operator** form:

$$y[k] = G(q^{-1})u[k] + H(q^{-1})e[k],$$

where $G(q^{-1})$ and $H(q^{-1})$ are known as plant and noise models, respectively.

Final Model

The last OE model in the TF form is written as:

$$y[k] = \frac{0.4621q^{-1}}{1 - 0.8826q^{-1}}u[k] + e[k]$$

Reference

1. *Principles of System Identification*, Arun K. Tangirala, CRC Press