DTU

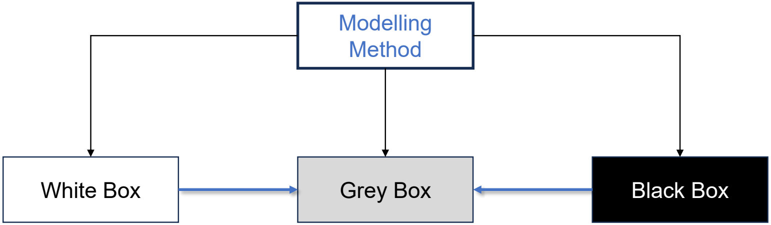
Industrial IoT for Digitization of Electronic Assets

System Identification and Parameters Estimation

Agenda

- White, Grey, and Black Box Models
- LTI system and properties
- Autoregressive Model in System Identification
- Autoregressive with eXogenous
- Parameters Estimation
- ARX & ARMAX Models' Structure
- Validation and Residual Analysis
- Order Selection

Model Types



White Box Models

- The structure of the model is known and is developed from fundamental laws of physics, thermodynamics, and heat transfer;
- The model parameters are well knownand used as inputs to the model.

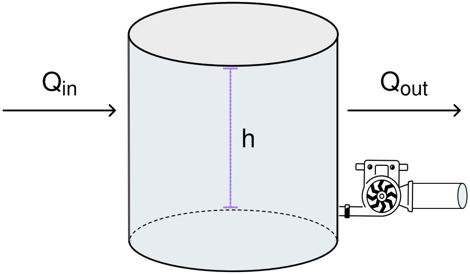
Grey Box Models

- Combination of black box and white box models.
- When data-driven models (black box) are partly supported by explicit models (equations) with a physical meaning.

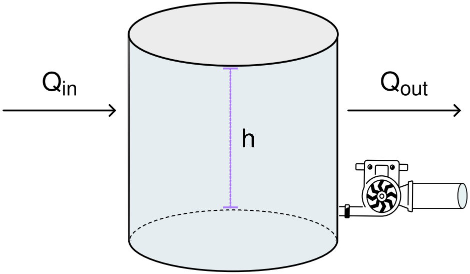
Black Box Models

- Purely data-driven/statistical/empirical models
- Uses mathematical equations from statistics to map influential inputs to the outputs.
- The source of data: on-board monitored data from sensors, data simulated from simulation tools, surveyor standard data from public benchmark datasets.

Example of White Box Model: Tank Model



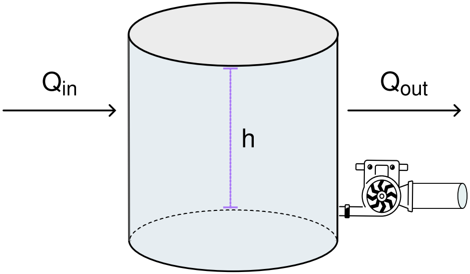
Example of White Box Model: Tank Model



Mass Conservation Law

$$\frac{\mathrm{d}v_t}{\mathrm{d}t} = q_t^{\text{in}} - q_t^{\text{out}} \tag{1}$$

Example of White Box Model: Tank Model

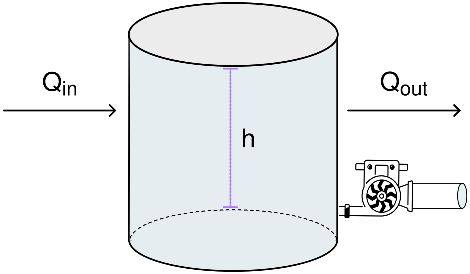


Mass Conservation Law

$$\frac{\mathrm{d}v_t}{\mathrm{d}t} = q_t^{\text{in}} - q_t^{\text{out}} \tag{1}$$

$$v_t = A \cdot h_t \tag{2}$$

Example of White Box Model: Tank Model



Mass Conservation Law

$$\frac{\mathrm{d}v_t}{\mathrm{d}t} = q_t^{\text{in}} - q_t^{\text{out}} \tag{1}$$

$$v_t = A \cdot h_t \tag{2}$$

$$\frac{\mathrm{d}h_t}{\mathrm{d}t} = \frac{1}{A}(q_t^{\text{in}} - q_t^{\text{out}}) \tag{3}$$

Example of Grey Box Model

Stochastic Diff. Equations

$$dx_t = f(x_t, u_t, t, \theta) + \sigma(u_t, t, \theta)dw$$
$$y_k = h(x_k, u_k, t_k, \theta) + e_k$$

- Continuous-Time Models
- Suitable for both linear and non-linear models.
- Used to represent systems influenced by both deterministic and stochastic components.

Example of Grey Box Model

Stochastic Diff. Equations

$$dx_t = f(x_t, u_t, t, \theta) + \sigma(u_t, t, \theta)dw$$
$$y_k = h(x_k, u_k, t_k, \theta) + e_k$$

- Continuous-Time Models
- Suitable for both linear and non-linear models.
- Used to represent systems influenced by both deterministic and stochastic components.

PINNs: Physics-informed neural networks

- A class of machine learning techniques that combines data-driven neural networks with physical equations.
- Minimizing the discrepancy a loss function formed by data and equations.

Example of Black Box Models

A black box model is a system or algorithm that makes predictions or decisions based solely on input and output data, without an interpretable framework that can explain the connection between the inputs and the outputs.

Assumptions

- **Linear Structure**
  - The system is described by linear difference equations
  - Input-output relationship is linear
  - Superposition principle

Assumptions

- **Linear Structure**
  - The system is described by linear difference equations
  - Input-output relationship is linear
  - Superposition principle
- **Time-Invariant Parameters**
  - Time-Invariant Parameters
  - Model coefficients remain constant over time

Assumptions

- **Linear Structure**
  - The system is described by linear difference equations
  - Input-output relationship is linear
  - Superposition principle
- **Time-Invariant Parameters**
  - Time-Invariant Parameters
  - Model coefficients remain constant over time
- **Noise Properties ε**
  - ***i.i.d***: Independent and identically distributed.
  - No correlation between noise values, i.e. past values do not influence future values.

Assumptions

- **Linear Structure**
  - The system is described by linear difference equations
  - Input-output relationship is linear
  - Superposition principle
- **Time-Invariant Parameters**
  - Time-Invariant Parameters
  - Model coefficients remain constant over time
- **Noise Properties ε**
  - ***i.i.d***: Independent and identically distributed.
  - No correlation between noise values, i.e. past values do not influence future values.

We assume the principle *"let the data speak for itself"*, i.e. we do not consider an explicit physical model but a class of autoregressive models.

Linear Time-Invariant (LTI) Systems

Definition:

A *Linear Time-Invariant* (LTI) system is a system where the principles of superposition (linearity) and shift invariance (time invariance) hold.

Linear Time-Invariant (LTI) Systems

Definition:

A *Linear Time-Invariant* (LTI) system is a system where the principles of superposition (linearity) and shift invariance (time invariance) hold.

Properties:

- **Linearity:** Given $y_1(t)$ and $y_2(t)$ the outputs corresponding to inputs $u_1(t)$ and $u_2(t)$, then for any constants a and b, the system's response to $au_1(t) + bu_2(t)$ is $ay_1(t) + by_2(t)$.
- **Time Invariance:** If the output of the system to an input $u(t)$ is $y(t)$, then the output to an input $u(t - t_0)$, for any time shift t_0, is $y(t - t_0)$.

Autoregressive Model (AR)

First Order Model

Let's consider the simplest form of an AR model:

$$y(t) = a_1 y(t-1) + \varepsilon_t$$

Where $\boldsymbol{y}(\boldsymbol{t}-1)$ is the *lag* of $\boldsymbol{y}$ at time $t-1$ and ϵ_t an error term.

n_a order model

$$y(t) = a_1 y(t-1) + a_2 y(t-2) + \cdots + a_{n_a} y(t-n_a) + \varepsilon_t$$

Exercise

$$y(t) = 50 + 0.7 \cdot y(t-1) - 0.1y(t-2) + \varepsilon_t$$

Time	Sales	Month
$y(t-2)$	90	April
$y(t-1)$	100	May
$y(t)$	?	June

Autoregressive eXogenous Model (ARX)

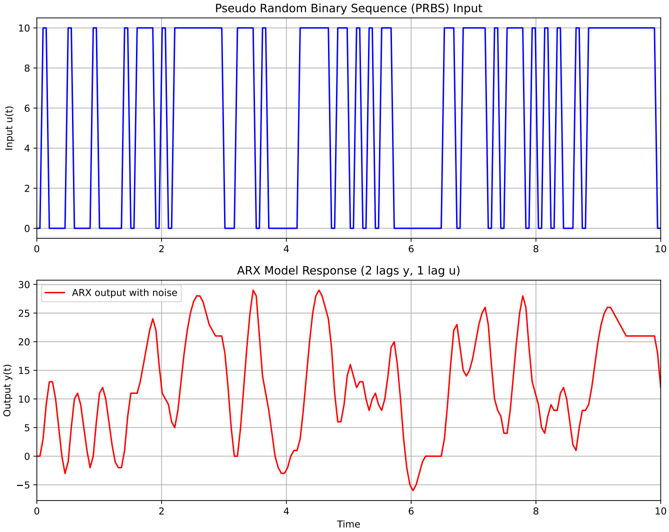
Given an **LTI** system, with $\boldsymbol{y}(\boldsymbol{t})$ the output signal *with exogenous* input $\boldsymbol{u}(\boldsymbol{t})$ with delay k, an ARX model can be defined as follows:

$$\begin{aligned} y(t) = {} & a_1 y(t-1) + \cdots + a_{n_a} y(t-n_a) \\ & + b_1 u(t-n_k) + \cdots + b_{n_b} u(t-n_k-n_{n_b}+1) + \varepsilon_t \end{aligned}$$

- $\boldsymbol{n_a}$: numbers of lags of the output signal y.
- $\boldsymbol{n_b}$: numbers of lags of input signal u.
- $\boldsymbol{n_k}$: delay order between the y and u.

$$\boxed{y(t) = \varphi(t)^T \theta + \varepsilon(t)}$$

Example



Pseudo Random Binary Sequence (PRBS) Input
Input u(t)
10
8
6
4
2
0
0
2
4
6
8
10
ARX Model Response (2 lags y, 1 lag u)
ARX output with noise
Output y(t)
30
25
20
15
10
5
0
−5
0
2
4
6
8
10
Time

AutoRegressive eXogenous Model (ARX)

Or equivalently:

$$\boxed{y(t) = \varphi^T(t)\theta + \varepsilon(t)}$$

$$\theta = [a_1, \ldots, a_{n_a}\ b_1, \ldots, b_{n_b}] \quad \textit{unknown parameters}$$

$$\varphi_t = [y_t, \ldots, y_{t-n_a},\ u_{t-n_k} \ldots, u_{t-n_k-n_b+1}]^T \quad \textit{regressors}$$

$$\varepsilon_t \sim N(0, \sigma^2) \quad \textit{error}$$

Remark

We denote $\hat{\boldsymbol{y}}(\boldsymbol{t}|\theta) = \varphi^T(\boldsymbol{t})\theta$ the estimated output prediction to emphasize that the estimate of $\hat{y}$ depends on the past data φ and the parameters vector θ with $\varepsilon = 0$ (best case scenario - **no residual error**).

Remark

We denote $\hat{\boldsymbol{y}}(\boldsymbol{t}|\theta) = \varphi^T(\boldsymbol{t})\theta$ the estimated output prediction to emphasize that the estimate of $\hat{y}$ depends on the past data φ and the parameters vector θ with $\varepsilon = 0$ (best case scenario - **no residual error**).

Remark

We denote $\hat{\boldsymbol{y}}(\boldsymbol{t}|\theta) = \varphi^T(\boldsymbol{t})\theta$ the estimated output prediction to emphasize that the estimate of $\hat{y}$ depends on the past data φ and the parameters vector θ with $\varepsilon = 0$ (best case scenario - **no residual error**).

- $\varphi(t)^T \rightarrow$ (past data)
- $\epsilon(\boldsymbol{t}) = 0 \rightarrow$ (known distribution)

What is still missing?

Remark

We denote $\hat{\boldsymbol{y}}(\boldsymbol{t}|\theta) = \varphi^T(\boldsymbol{t})\theta$ the estimated output prediction to emphasize that the estimate of $\hat{y}$ depends on the past data φ and the parameters vector θ with $\varepsilon = 0$ (best case scenario - **no residual error**).

- $\varphi(t)^T \rightarrow$ (past data)
- $\epsilon(\boldsymbol{t}) = 0 \rightarrow$ (known distribution)

What is still missing?

Remark

We denote $\hat{\boldsymbol{y}}_{t|\theta} = \varphi^T \boldsymbol{\theta}$ the estimated output prediction to emphasize that the estimate of $\hat{y}$ depends from the past data on the parameters vector θ with $\varepsilon = 0$ (best case scenario - **no residual error**).

- $\varphi^T \rightarrow$ regressors vector (past data)
- $\bar{\epsilon} = 0 \rightarrow$ (known distribution)
- $\boldsymbol{\theta}$ is the coefficients vector (to be estimated)

Parameters Estimation

We don't know θ yet, but we have collected a set Z^N of measured data.

$$Z^N = \{u(-n), y(-n) \dots u(N-1), y(N-1)\},\ n = max\{n_a, n_b + n_k - 1\}$$

Therefore, we use the least squared error to find the optimal parameters vector θ^* that minimize the prediction error between $\hat{y}_{t+1|\theta}$ and y_t, namely:

$$\theta^* = \arg\min_{\theta}\{\mathscr{L}(\theta, Z^N)\}$$

$$\mathscr{L}(\theta, Z^N) = \sum_{t=0}^{N-1}(y(t) - \hat{y}(t|\theta))^2 = \sum_{t=0}^{N-1}(y(t) - \varphi(t)^T\theta)^2$$

Parameters Estimation

$\mathscr{L}(\theta, Z^N)$ is a quadratic function of θ. Therefore, θ^* can be found by zeroing the derivative of $\mathscr{L}$.

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathscr{L}_N(\theta, Z^N) = 2\sum_{t=0}^{N-1} \varphi(t)(y(t) - \varphi(t)^T\theta) = 0$$

Or:

$$\sum_{t=0}^{N-1} \varphi(t)y(t) = \sum_{t=0}^{N-1} \varphi(t)\varphi(t)^T\theta$$

Parameters Estimation

Or:

$$\sum_{t=0}^{N-1} \varphi(t)y(t) = \sum_{t=0}^{N-1} \varphi(t)\varphi(t)^T\theta$$

Finally, θ^* can be derived:

$$\boxed{\theta^* = \left[\sum_{t=0}^{N-1} \varphi(t)\varphi(t)^T\right]^{-1} \left[\sum_{k=0}^{N-1} \varphi(t)y(t)\right]}$$

How to calculate the error of a fitted model?

MAE (Mean Absolute Error)

The MAE is the average of the absolute errors between prediction $\hat{y}_i$ and actual observations y_i. It's calculated as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

RMSE (Root Mean Square Error)

The RMSE is the square root of the average of squared differences between prediction $\hat{y}_i$ and actual observation y_i. It's calculated as:

$$\text{RMSE} = \sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

RRMSE (Relative Root Mean Square Error)

RRMSE is the RMSE normalized by a characteristic scale of the actual observed values.$\hat{y}_i$.

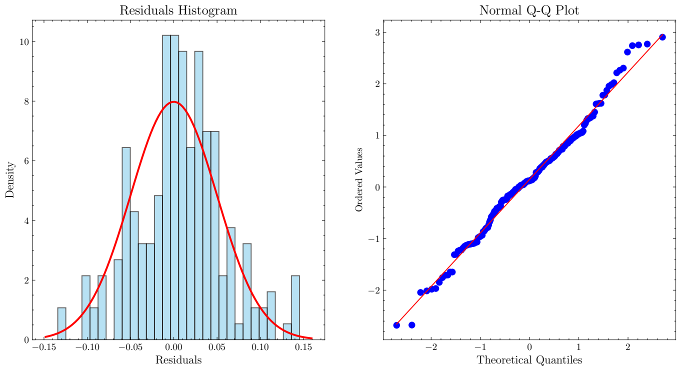
$$\text{RMSE} = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(\hat{y})^2}}$$

Do you remember the assumption on $\varepsilon(t)$?

We previously assumed that the residual of the model, namely $(y - \hat{y})$ are distributed as $\varepsilon(t) \sim \mathcal{N}(0, \sigma^2)$.

Model Validation

Shapiro-Wilk test confirmed that the residuals are white noise and the Q-Q Plot below shows that the distribution of the residual is compatible with the noise introduced in our data $\sim \mathcal{N}(0, \sigma^2 = 0.1)$.



State Space Representation of an ARX Model

- Consider a dynamic system with two lags in output and one in input:

$$y(k) = -a_1 y(k-1) - a_2 y(k-2) + b_0 u(k) + b_1 u(k-1)$$

- This can be written in state space form by choosing states:

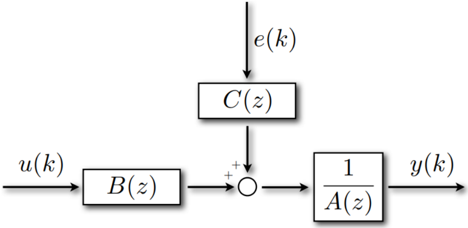
$$x_1(k) = y(k), \quad x_2(k) = y(k-1)$$

- The resulting state space model is:

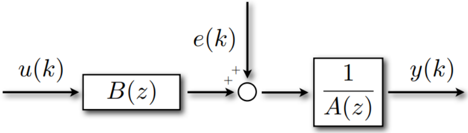
$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \end{bmatrix} = \underbrace{\begin{bmatrix} -a_1 & -a_2 \\ 1 & 0 \end{bmatrix}}_{\mathbf{A}} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} + \underbrace{\begin{bmatrix} b_0 & b_1 \\ 0 & 0 \end{bmatrix}}_{\mathbf{B}} \begin{bmatrix} u(k) \\ u(k-1) \end{bmatrix}$$

$$y(k) = \underbrace{\begin{bmatrix} 1 & 0 \end{bmatrix}}_{\mathbf{C}} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} + \underbrace{0}_{\mathbf{D}} u(k)$$

What if the noise is colored?

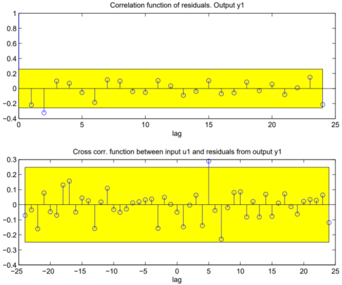


ARMAX model

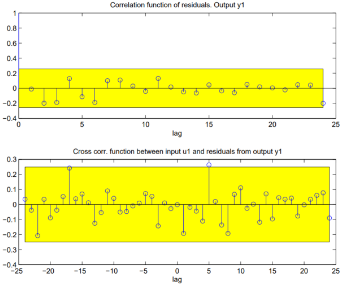


ARX model

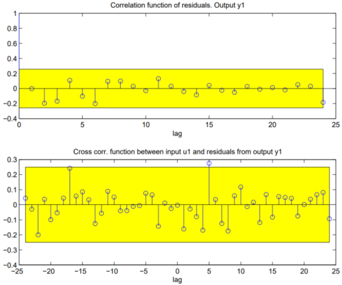
Model Complexity



ARX(2,2,1)



ARX(3,3,1)



ARX(4,4,1)

Model Selection: *AIC* & *BIC*

AIC and BIC are fundamental statistical tools used to select the best model from a group of potential models based on their performance with a given dataset. They are particularly crucial when there are multiple models, as they help balance fitting the data and model complexity, thus preventing overfitting or underfitting.

Model Selection: *AIC* & *BIC*

Akaike Information Criterion (AIC)

Definition:

$$\boxed{\mathrm{AIC} = 2k - 2\ln(L)}$$

where:

- **k** is the number of parameters
- **L** is the maximum likelihood of the model.

Bayesian Information Criterion (BIC)

Definition:

$$\boxed{\mathrm{BIC} = \ln(n)k - 2\ln(L)}$$

where:

- **n** is the number of observations
- **k** is the number of parameters
- **L** is the maximum likelihood of the model.

Model Selection: *AIC* & *BIC*

Akaike Information Criterion (AIC)

Definition:

- Balances model fit and complexity.
- More emphasis on goodness of fit than on simplicity.

Limitations:

- Relative measure; cannot be used to test a single model.
- Can be biased in small samples.

Bayesian Information Criterion (BIC)

Definition:

- Includes a penalty term for the number of observations
- making it more reliable for larger datasets.
- Favors simpler models compared to AIC.
- Assumes that the model errors are independently and identically distributed.

Building a Transfer Function from ARX and ARMAX Models

ARX Model:

- ARX model equation:

$$A(q)y(t) = B(q)u(t) + e(t)$$

where:
  - $A(q) = 1 + a_1 q^{-1} + \cdots + a_n q^{-n}$
  - $B(q) = b_1 q^{-1} + \cdots + b_m q^{-m}$
- Transfer Function:

$$G(q) = \frac{B(q)}{A(q)}$$

ARMAX Model:

- ARMAX model equation:

$$A(q)y(t) = B(q)u(t) + C(q)e(t)$$

where:
  - $C(q) = 1 + c_1 q^{-1} + \cdots + c_p q^{-p}$
- Transfer Function:

$$G(q) = \frac{B(q)}{A(q)}, \quad H(q) = \frac{1}{C(q)}$$

where $H(q)$ models noise dynamics.