SYSTEM IDENTIFICATION

Michele TARAGNA

Dipartimento di Automatica e Informatica

Politecnico di Torino

michele.taragna@polito.it

II level Specializing Master in Automatica and Control Technologies

Class **“System Identification, Estimation and Filtering”**

Academic Year 2010/2011

System identification

- System identification is aimed at constructing or selecting mathematical models $\mathcal{M}$ of dynamical data generating systems $\mathcal{S}$ to serve certain purposes (forecast, diagnostic, control, etc.)
- A first step is to determine a class $\mathfrak{M}$ of models within which the search for the most suitable model is to be conducted
- Classes of parametric models $\mathcal{M}(\theta)$ are often considered, where the parameter vector θ belongs to some set Θ, i.e., $\mathfrak{M} = \{\mathcal{M}(\theta) : \theta \in \Theta\}$

  $$\Downarrow$$

  the choice problem is tackled as a parametric estimation problem
- We start by discussing two model classes for linear time-invariant (LTI) systems:
  - transfer-function models
  - state-space models

Transfer-function models

- The transfer-function models, known also as black-box or Box-Jenkins models, involve external variables only (i.e., input and output variables) and do not require any auxiliary variable
- Different structures of transfer-function models are available:
  - equation error or ARX model structure
  - ARMAX model structure
  - output error (OE) model structure

Equation error or ARX model structure

- The input-output relationship is a linear difference equation:
  $y(t)+a_1y(t-1)+a_2y(t-2)+\cdots+a_{n_a}y(t-n_a)=b_1u(t-1)+\cdots+b_{n_b}u(t-n_b)+e(t)$
  where the white-noise term $e(t)$ enters as a direct error
- Let us denote by z^{-1} the unitary delay operator, such that $z^{-1}y(t) = y(t-1)$, $z^{-2}y(t) = y(t-2)$, etc., and introduce the polynomials:

$$A(z) = 1 + a_1z^{-1} + a_2z^{-2} + \cdots + a_{n_a}z^{-n_a}$$
$$B(z) = b_1z^{-1} + b_2z^{-2} + \cdots + b_{n_b}z^{-n_b}$$

then, the above input-output relationship can be written as:

$$A(z)y(t) = B(z)u(t) + e(t) \qquad \Rightarrow$$
$$y(t) = \frac{B(z)}{A(z)}u(t) + \frac{1}{A(z)}e(t) = G(z)u(t) + H(z)e(t)$$

where

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

- If the input $u(\cdot)$ is present, also known as exogenous variable, then the model:
$$A(z)y(t) = B(z)u(t) + e(t)$$
contains the *autoregressive (AR)* $A(z)y(t)$ and the *exogenous (X)* $B(z)u(t)$ parts.
The integers n_a and n_b are the *orders* of these two parts of the *ARX* model, denoted as $ARX(n_a, n_b)$
- If $n_a = 0$, then $A(z) = 1$ and $y(t)$ is modeled as a *finite impulse response (FIR)*
- If the input $u(\cdot)$ is missing, then the model:
$$A(z)y(t) = e(t)$$
contains only the *autoregressive (AR)* $A(z)y(t)$ part.
The integer n_a is the *order* of the resulting *AR* model, denoted as $AR(n_a)$

ARMAX model structure

- The input-output relationship is a linear difference equation:

$$y(t) + a_1 y(t-1) + a_2 y(t-2) + \cdots + a_{n_a} y(t-n_a) =$$
$$= b_1 u(t-1) + \cdots + b_{n_b} u(t-n_b) + e(t) + c_1 e(t-1) + \cdots + c_{n_c} e(t-n_c)$$

  where the white-noise term $e(t)$ enters as a linear combination of n_c+1 samples

- By introducing the polynomials:

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_{n_a} z^{-n_a}$$
$$B(z) = b_1 z^{-1} + b_2 z^{-2} + \cdots + b_{n_b} z^{-n_b}$$
$$C(z) = 1 + c_1 z^{-1} + c_2 z^{-2} + \cdots + c_{n_c} z^{-n_c}$$

  the above input-output relationship can be written as:

$$A(z)y(t) = B(z)u(t) + C(z)e(t) \qquad \Rightarrow$$
$$y(t) = \frac{B(z)}{A(z)} u(t) + \frac{C(z)}{A(z)} e(t) = G(z)u(t) + H(z)e(t)$$

  where

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

- If the exogenous variable $u(\cdot)$ is present, then the model:
$$A(z)y(t) = B(z)u(t) + C(z)e(t)$$
contains the *autoregressive (AR)* part $A(z)y(t)$, the *exogenous (X)* part $B(z)u(t)$ and the *moving average (MA)* part $C(z)e(t)$, which is a colored noise instead of a white one.
The integers n_a, n_b and n_c are the orders of these three parts of the *ARMAX* model, denoted as $ARMAX(n_a, n_b, n_c)$
- If the input $u(\cdot)$ is missing, then the model:
$$A(z)y(t) = C(z)e(t)$$
contains only the *autoregressive* $A(z)y(t)$ and the *moving average* $C(z)e(t)$ parts.
The integers n_a and n_c are the orders of the resulting *ARMA* model, denoted as $ARMA(n_a, n_c)$

Output error or OE model structures

- The relationship between input and undisturbed output is a linear difference equation:

$$w(t) + f_1 w(t-1) + \cdots + f_{n_f} w(t-n_f) = b_1 u(t-1) + \cdots + b_{n_b} u(t-n_b)$$

and the model output is corrupted by white measurement noise:

$$y(t) = w(t) + e(t)$$

- By introducing the polynomials:

$$F(z) = 1 + f_1 z^{-1} + f_2 z^{-2} + \cdots + f_{n_f} z^{-n_f}$$

$$B(z) = b_1 z^{-1} + b_2 z^{-2} + \cdots + b_{n_b} z^{-n_b}$$

the above input-undisturbed output relationship can be written as:

$$F(z)w(t) = B(z)u(t) \qquad \Rightarrow$$

$$y(t) = w(t) + e(t) = \frac{B(z)}{F(z)} u(t) + e(t) = G(z)u(t) + e(t)$$

where

$$G(z) = \frac{B(z)}{F(z)}$$

- The integers n_b and n_f are the orders of the *OE* model, denoted as $OE(n_b, n_f)$

State-space models

The discrete-time, linear time-invariant model $\mathcal{M}$ is described by:

$$\mathcal{M}: \begin{cases} x(t+1) = Ax(t) + Bu(t) + v_1(t) \\ y(t) = Cx(t) + v_2(t) \end{cases} \qquad t = 1, 2, \ldots$$

where:

- $x(t) \in \mathbb{R}^n$, $y(t) \in \mathbb{R}^q$, $u(t) \in \mathbb{R}^p$, $v_1(t) \in \mathbb{R}^n$, $v_2(t) \in \mathbb{R}^q$
- the process noise $v_1(t)$ and the measurement noise $v_2(t)$ are uncorrelated white noises with zero mean value, i.e.:
  $v_1(t) \sim WN(0, V_1)$ with $V_1 \in \mathbb{R}^{n \times n}$, $v_2(t) \sim WN(0, V_2)$ with $V_2 \in \mathbb{R}^{q \times q}$
- $A \in \mathbb{R}^{n \times n}$ is the state matrix, $B \in \mathbb{R}^{n \times p}$ is the input matrix,
  $C \in \mathbb{R}^{q \times n}$ is the output matrix

The transfer matrix between the exogenous input u and the output y is:

$$G(z) = C\left(zI_n - A\right)^{-1} B$$

The system identification procedure

The system identification problem may be solved using an iterative approach:

1. Collect the data set
   - If possible, design the experiment so that the data become maximally informative
   - If useful and/or necessary, apply some prefiltering technique of the data
2. Choose the model set or the model structure, so that it is suitable for the aims
   - A *physical* model with some unknown parameters may be constructed by exploiting the possible a priori knowledge and insight
   - Otherwise, a *black box* model may be employed, whose parameters are simply tuned to fit the data, without reference to the physical background
   - Otherwise, a *gray box* model may be used, with adjustable parameters having physical interpretation
3. Determine the suitable complexity level of the model set or model structure
4. Tune the parameters to pick the “best” model in the set, guided by the data
5. Perform a model validation test: if the model is OK, then use it, otherwise revise it

The predictive approach

Let us consider a class $\mathfrak{M}$ of parametric models $\mathcal{M}(\theta)$:

$$\mathfrak{M} = \{\mathcal{M}(\theta) : \theta \in \Theta\}$$

where the parameter vector θ belongs to some set Θ

The data are the measurements collected at the time instants t from 1 to N

- of the variable $y(t)$, in the case of time series
- of the input $u(t)$ and the output $y(t)$, in the case of input-output systems

Given a model $\mathcal{M}(\theta)$, a corresponding predictor $\hat{\mathcal{M}}(\theta)$ can be associated that provides the optimal one-step prediction $\hat{y}(t+1|t)$ of $y(t+1)$ on the basis of the data, i.e.,

- in the case of time series, the predictor is given by:

$$\hat{\mathcal{M}}(\theta) : \hat{y}(t+1) = \hat{y}(t+1|t) = f\left(y^t, \theta\right)$$

- in the case of input-output systems, the predictor is given by:

$$\hat{\mathcal{M}}(\theta) : \hat{y}(t+1) = \hat{y}(t+1|t) = f\left(u^t, y^t, \theta\right)$$

with $y^t = \{y(t), y(t-1), y(t-2), \ldots, y(1)\}, u^t = \{u(t), u(t-1), u(t-2), \ldots, u(1)\}$

Given a model $\mathcal{M}(\theta)$ with a fixed value of the parameter vector θ, the prediction error at the time instant $t+1$ is given by:

$$\varepsilon(t+1) = y(t+1) - \hat{y}(t+1|t)$$

and the overall mean-square error (MSE) is defined as:

$$J_N(\theta) = \frac{1}{N}\sum_{t=\tau}^{N} \varepsilon(t)^2$$

where τ is the first time instant at which the prediction $\hat{y}(\tau|\tau-1)$ of $y(\tau)$ can be computed from the data ($\tau = 1$ is often assumed)

In the predictive approach to system identification, the parameters of the model $\mathcal{M}(\theta)$ in the class $\mathfrak{M}$ are tuned to minimize the criterion $J_N(\theta)$ over all $\theta \in \Theta$, i.e.,

$$\hat{\theta}_N = \arg\min_{\theta\in\Theta} J_N(\theta)$$

If the model quality is high, the prediction error has to be white, i.e., without its own dynamics, since the dynamics contained in the data has to be explained by the model

$\Rightarrow$ many different *whiteness tests* can be performed on the sequence $\varepsilon(t)$

Models in predictor form

Let us consider the transfer-function model

$$\mathcal{M}(\theta) : y(t) = G(z)u(t) + H(z)e(t)$$

where $e(t)$ is a white noise with zero mean value

The term $v(t) = H(z)e(t)$ is called *residual* and has to be small, so that the model $\mathcal{M}(\theta)$ could satisfactorily describe the input-output relationship of a given system $\mathcal{S}$

$$\Downarrow$$

It is typically assumed that $v(t)$ is a stationary process, i.e., a sequence of random variables whose joint probability distribution does not change over time or space $\Rightarrow$ the following assumptions can be made, leading to the canonical representation of $v(t)$:

1. $H(z)$ is the ratio of two polynomials with the same degree that are:
   - monic, i.e., such that the coefficients of the highest order terms are equal to 1
   - coprime, i.e., without common roots
2. both the numerator and the denominator of $H(z)$ are asymptotically stable, i.e., the magnitude of all the zeros and poles of $H(z)$ is less than 1

The predictor associated to $\mathcal{M}(\theta)$ can be derived from the model equation as follows:

1. subtract $y(t)$ from both sides: $0 = -y(t) + G(z)u(t) + H(z)e(t)$
2. divide by $H(z)$: $0 = -\frac{1}{H(z)}y(t) + \frac{G(z)}{H(z)}u(t) + e(t)$
3. add $y(t)$ to both sides: $y(t) = \left[1 - \frac{1}{H(z)}\right] y(t) + \frac{G(z)}{H(z)}u(t) + e(t)$

Since $H(z)$ is the ratio of two monic polynomials with the same degree, then:

$$\frac{1}{H(z)} = 1 + \alpha_1 z^{-1} + \alpha_2 z^{-2} + \ldots \Rightarrow 1 - \frac{1}{H(z)} = -\alpha_1 z^{-1} - \alpha_2 z^{-2} - \ldots \Rightarrow$$

$$\left[1 - \frac{1}{H(z)}\right] y(t) = -\alpha_1 y(t-1) - \alpha_2 y(t-2) - \ldots = f_y(y^{t-1})$$

with $y^{t-1} = \{y(t-1), y(t-2), \ldots\}$. Analogously, since $G(z)$ is strictly proper:

$$\frac{G(z)}{H(z)} = G(z)\left(1 + \alpha_1 z^{-1} + \alpha_2 z^{-2} + \ldots\right) = \beta_1 z^{-1} + \beta_2 z^{-2} + \ldots \Rightarrow$$

$$\frac{G(z)}{H(z)}u(t) = \beta_1 u(t-1) + \beta_2 u(t-2) + \ldots = f_u(u^{t-1})$$

with $u^{t-1} = \{u(t-1), u(t-2), \ldots\}$ and then:

$$y(t) = f_y(y^{t-1}) + f_u(u^{t-1}) + e(t)$$

In the model equation

$$y(t) = f_y(y^{t-1}) + f_u(u^{t-1}) + e(t)$$

the output $y(t)$ depends on past values u^{t-1} and y^{t-1} of the input and the output, while the white noise term $e(t)$ is unpredictable and independent of u^{t-1} and y^{t-1}

$$\Downarrow$$

the best prediction of $e(t)$ is provided by its mean value, which is equal to 0

$$\Downarrow$$

the optimal one-step predictor of the model $\mathcal{M}(\theta)$ is given by:

$$\hat{\mathcal{M}}(\theta) : \hat{y}(t) = \hat{y}(t|t-1) = \left[1 - \frac{1}{H(z)}\right] y(t) + \frac{G(z)}{H(z)} u(t)$$

ARX, AR and FIR models in predictor form

In the case of the *ARX* transfer-function model:

$$\mathcal{M}(\theta) : y(t) = G(z)u(t) + H(z)e(t), \quad \text{with} \quad G(z) = \frac{B(z)}{A(z)},\ H(z) = \frac{1}{A(z)}$$

the optimal predictor is given by:

$$\hat{\mathcal{M}}(\theta) : \hat{y}(t) = [1 - A(z)]\, y(t) + B(z)u(t)$$

- $\hat{y}(t)$ is a linear combination of past values of the input and the output, independent of past predictions
- $\hat{y}(t)$ is linear in the parameters a_i and b_i of the polynomials $A(z)$ and $B(z)$
- the predictor is stable for any value of the parameters that define $A(z)$ and $B(z)$

In the case of the AR transfer-function model, where $B(z) = 0$, then:

$$\hat{\mathcal{M}}(\theta) : \hat{y}(t) = [1 - A(z)]\, y(t)$$

In the case of the FIR transfer-function model, where $A(z) = 1$, then:

$$\hat{\mathcal{M}}(\theta) : \hat{y}(t) = B(z)u(t)$$

ARMAX, ARMA and MA models in predictor form

In the case of the *ARMAX* transfer-function model:

$$\mathcal{M}(\theta) : y(t) = G(z)u(t) + H(z)e(t), \quad \text{with} \quad G(z) = \frac{B(z)}{A(z)}, \; H(z) = \frac{C(z)}{A(z)}$$

the optimal predictor is given by:

$$\hat{\mathcal{M}}(\theta) : \hat{y}(t) = \left[1 - \frac{A(z)}{C(z)}\right] y(t) + \frac{B(z)}{C(z)} u(t)$$

- $\hat{y}(t)$ is nonlinear in the parameters a_i, b_i, c_i of the polynomials $A(z), B(z), C(z)$
- the predictor stability depends on the values of the parameters that define $C(z)$

In the case of the *ARMA* transfer-function model, where $B(z) = 0$, then:

$$\hat{\mathcal{M}}(\theta) : \hat{y}(t) = \left[1 - \frac{A(z)}{C(z)}\right] y(t)$$

In the case of the *MA* transfer-function model, where $B(z) = 0$ and $A(z) = 1$, then:

$$\hat{\mathcal{M}}(\theta) : \hat{y}(t) = \left[1 - \frac{1}{C(z)}\right] y(t)$$

OE models in predictor form

In the case of the *OE* transfer-function model:

$$\mathcal{M}(\theta) : y(t) = G(z)u(t) + H(z)e(t), \quad \text{with} \quad G(z) = \frac{B(z)}{F(z)},\ H(z) = 1$$

the optimal predictor is given by:

$$\hat{\mathcal{M}}(\theta) : \hat{y}(t) = \frac{B(z)}{F(z)}u(t)$$

- $\hat{y}(t)$ is a linear combination of past values of the exogenous input only, independent of past predictions
- $\hat{y}(t)$ is nonlinear in the parameters b_i, f_i of the polynomials $B(z)$ and $F(z)$
- the predictor stability depends on the values of the parameters that define $F(z)$

Asymptotic analysis of prediction-error identification methods

Using the prediction-error identification methods (PEM), the optimal model in the parametric class $\mathfrak{M} = \{\mathcal{M}(\theta) : \theta \in \Theta\}$ is obtained by minimizing the "size" of the prediction-error sequence $\varepsilon(\cdot)$, i.e.:

$$J_N(\theta) = \frac{1}{N}\sum_{t=\tau}^{N}\varepsilon(t)^2 \text{ or, in general, } J_N(\theta) = \frac{1}{N}\sum_{t=\tau}^{N}\ell\left(\varepsilon(t)\right)$$

where $\ell\left(\cdot\right)$ is a scalar-valued (typically positive) function

Goal: analyze the asymptotic (i.e., as $N \to \infty$) characteristics of the optimal estimate

$$\hat{\theta}_N = \arg\min_{\theta\in\Theta} J_N(\theta)$$

Assumptions: the predictor form $\hat{\mathcal{M}}(\theta)$ of the model $\mathcal{M}(\theta)$ is stable and the sequences $u(\cdot)$ and $y(\cdot)$ are stationary processes $\Rightarrow$ the one-step prediction $\hat{y}(\cdot)$ and the prediction error $\varepsilon(\cdot) = y(\cdot) - \hat{y}(\cdot)$ are stationary processes as well $\Rightarrow$

$$J_N(\theta) = \frac{1}{N}\sum_{t=\tau}^{N}\varepsilon(t)^2 \xrightarrow[N\to\infty]{} \bar{J}(\theta) = E\left[\varepsilon(t)^2\right]$$

Let us denote by $\mathcal{D}_\Theta$ the set of minimum points of $\bar{J}(\theta)$, i.e.:

$$\mathcal{D}_\Theta = \left\{\bar{\theta} : \bar{J}(\bar{\theta}) \leq \bar{J}(\theta), \forall \theta \in \Theta\right\}$$

Result #1:

if the data generating system $\mathcal{S} \in \mathfrak{M}$, i.e., $\exists \theta_o \in \Theta : \mathcal{S} = \mathcal{M}(\theta_o) \Rightarrow \theta_o \in \mathcal{D}_\Theta$

Result #2:

1. if $\mathcal{S} \in \mathfrak{M}$ and $\mathcal{D}_\Theta = \{\theta_o\}$ (i.e., $\mathcal{D}_\Theta$ is a singleton) $\Rightarrow \hat{\theta}_N \xrightarrow[N \to \infty]{} \theta_o$
2. if $\mathcal{S} \in \mathfrak{M}$ and $\exists \bar{\theta} \in \mathcal{D}_\Theta : \bar{\theta} \neq \theta_o$ (i.e., $\mathcal{D}_\Theta$ is not a singleton) $\Rightarrow$ asymptotically:
   - either $\hat{\theta}_N$ tends to a point in $\mathcal{D}_\Theta$ (not necessarily θ_o)
   - or it does not converge to any particular point of $\mathcal{D}_\Theta$ but wanders around in $\mathcal{D}_\Theta$
3. if $\mathcal{S} \notin \mathfrak{M}$ and $\mathcal{D}_\Theta = \left\{\bar{\theta}\right\}$ (i.e., $\mathcal{D}_\Theta$ is a singleton) $\Rightarrow \hat{\theta}_N \xrightarrow[N \to \infty]{} \bar{\theta}$ and $\mathcal{M}(\bar{\theta})$ is the best approximation of $\mathcal{S}$ within $\mathfrak{M}$
4. if $\mathcal{S} \notin \mathfrak{M}$ and $\mathcal{D}_\Theta$ is not a singleton $\Rightarrow$ asymptotically, either $\hat{\theta}_N$ tends to a point in $\mathcal{D}_\Theta$ or it wanders around in $\mathcal{D}_\Theta$

To measure the uncertainty and the convergence rate of the estimate $\hat{\theta}_N$, we have to study the random variable $\hat{\theta}_N - \bar{\theta}$, being $\bar{\theta}$ the limit of $\hat{\theta}_N$ as $N \to \infty$

Result #3:
if $\mathcal{S} \in \mathfrak{M}$ and $\mathcal{D}_\Theta = \{\theta_o\} \in \mathbb{R}^n$, then:

- $\hat{\theta}_N - \theta_o$ decays as $1/\sqrt{N}$ for $N \to \infty$
- the random variable $\sqrt{N}(\hat{\theta}_N - \theta_o)$ is asymptotically normally distributed:

$$\sqrt{N}(\hat{\theta}_N - \theta_o) \sim As\, \mathcal{N}\left(0, \bar{P}\right)$$

where

$$\bar{P} = Var[\varepsilon(t, \theta_o)]\bar{R}^{-1} \in \mathbb{R}^{n \times n} \qquad \text{(asymptotic variance matrix)}$$

$$\bar{R} = E\left[\psi(t, \theta_o)\psi(t, \theta_o)^T\right] \in \mathbb{R}^{n \times n}$$

$$\psi(t, \theta) = -\left[\frac{d}{d\theta}\varepsilon(t, \theta)\right]^T = -\left[\frac{d}{d\theta}\hat{y}(t, \theta)\right]^T \in \mathbb{R}^n$$

$$\Downarrow$$

$$\hat{\theta}_N \sim As\, \mathcal{N}\left(\theta_o, \frac{1}{N}\bar{P}\right)$$

Note that the asymptotic variance matrix $\bar{P}$ can be directly estimated from data as follows, having processed N data points and determined $\hat{\theta}_N$:

$$\bar{P} = Var[\varepsilon(t, \theta_o)]\bar{R}^{-1} \approx \hat{P}_N = \hat{\sigma}_N^2 \hat{R}_N^{-1}$$

$$\hat{\sigma}_N^2 = \frac{1}{N}\sum_{t=1}^{N} \varepsilon(t, \hat{\theta}_N)^2 \in \mathbb{R}$$

$$\hat{R}_N = \frac{1}{N}\sum_{t=1}^{N} \psi(t, \hat{\theta}_N)\psi(t, \hat{\theta}_N)^T \in \mathbb{R}^{n \times n}$$

$\Rightarrow$ the estimate uncertainty intervals can be derived from data

Linear regressions and least-squares method

In the case of equation error or *ARX* models, the optimal predictor is given by:

$$\hat{\mathcal{M}}(\theta) : \hat{y}(t) = [1 - A(z)]\, y(t) + B(z)u(t)$$

with $A(z) = 1 + a_1 z^{-1} + \cdots + a_{n_a} z^{-n_a}$, $B(z) = b_1 z^{-1} + \cdots + b_{n_b} z^{-n_b}$

$$\Downarrow$$

$$\begin{aligned}\hat{y}(t) &= \left(-a_1 z^{-1} - \cdots - a_{n_a} z^{-n_a}\right) y(t) + \left(b_1 z^{-1} + \cdots + b_{n_b} z^{-n_b}\right) u(t) = \\ &= -a_1 y(t-1) - \cdots - a_{n_a} y(t-n_a) + b_1 u(t-1) + \cdots + b_{n_b} u(t-n_b) = \\ &= \varphi(t)^T \theta = \hat{y}(t,\theta)\end{aligned}$$

where

$$\varphi(t) = [-y(t-1) \ \cdots \ -y(t-n_a) \ u(t-1) \ \cdots \ u(t-n_b)]^T \in \mathbb{R}^{n_a+n_b}$$

$$\theta = [a_1 \ \cdots \ a_{n_a} \ b_1 \ \cdots \ b_{n_b}]^T \in \mathbb{R}^{n_a+n_b}$$

i.e., it defines a *linear regression* $\Rightarrow$ the vector $\varphi(t)$ is known as the *regression vector*

Since the prediction error at the time instant t is given by:

$$\varepsilon(t,\theta) = y(t) - \hat{y}(t,\theta) = y(t) - \varphi(t)^T\theta, \quad t = 1,\ldots,N$$

and the optimality criterion (assuming $\tau = 1$, for the sake of simplicity) is quadratic:

$$J_N(\theta) = \frac{1}{N}\sum_{t=1}^{N}\varepsilon(t,\theta)^2$$

the optimal parameter vector $\hat{\theta}_N$ that minimizes $J_N(\theta)$ over all $\theta \in \Theta = \mathbb{R}^{n_a+n_b}$ is obtained by solving the normal equation system:

$$\left[\sum_{t=1}^{N}\varphi(t)\,\varphi(t)^T\right]\theta = \sum_{t=1}^{N}\varphi(t)\,y(t)$$

- if the matrix $\left[\sum_{t=1}^{N}\varphi(t)\,\varphi(t)^T\right]$ is nonsingular (known as *identifiability condition*), then there exists a single unique solution given by the *least-squares (LS) estimate*:

$$\hat{\theta}_N = \left[\sum_{t=1}^{N}\varphi(t)\,\varphi(t)^T\right]^{-1}\sum_{t=1}^{N}\varphi(t)\,y(t)$$

- otherwise, there are infinite solutions

Remark: the least-squares method can be applied to any model (not necessarily *ARX*) such that the corresponding predictor is a linear or affine function of θ:

$$\hat{y}(t,\theta) = \varphi(t)^T\theta + \mu(t)$$

where $\mu(t) \in \mathbb{R}$ is a known data-dependent vector. In fact, if the identifiability condition is satisfied, then: $\hat{\theta}_N = \left[\sum_{t=1}^{N} \varphi(t)\,\varphi(t)^T\right]^{-1} \sum_{t=1}^{N} \varphi(t)\,(y(t)-\mu(t))$

Such a situation may occur in many different situations:

- when some coefficients of the polynomials $A(z), B(z)$ of an *ARX* model are known
- when the predictor (even of a nonlinear model) can be written as a linear function of θ, by suitably choosing $\varphi(t)$

Example: given the nonlinear dynamic model

$$y(t) = ay(t-1)^2 + b_1u(t-3) + b_2u(t-5)^3 + e(t), \quad e(\cdot) \sim WN(0,\sigma^2)$$

the corresponding predictor is linear in the unknown parameters:

$$\hat{\mathcal{M}}(\theta) : \hat{y}(t) = ay(t-1)^2 + b_1u(t-3) + b_2u(t-5)^3 = \varphi(t)^T\theta$$

with $\varphi(t) = \begin{bmatrix} y(t-1)^2 & u(t-3) & u(t-5)^3 \end{bmatrix}^T$ and $\theta = \begin{bmatrix} a & b_1 & b_2 \end{bmatrix}^T$

Probabilistic analysis of the least-squares method

Let the predictor $\hat{\mathcal{M}}(\theta)$ of $\mathcal{M}(\theta)$ be stable and $u(\cdot)$, $y(\cdot)$ be stationary processes.

The least-squares method is a PEM method $\Rightarrow$ the previous asymptotic results hold

$\Rightarrow$ asymptotically, either $\hat{\theta}_N$ tends to a point in $\mathcal{D}_\Theta$ or wanders around in $\mathcal{D}_\Theta$, where $\mathcal{D}_\Theta = \left\{\bar{\theta} : \bar{J}(\bar{\theta}) \leq \bar{J}(\theta), \forall \theta \in \Theta\right\}$ is the set of minimum points of $\bar{J}(\theta) = E\left[\varepsilon(t)^2\right]$.

If $\mathcal{S} \in \mathfrak{M} \Rightarrow \exists \theta_o \in \mathcal{D}_\Theta : \mathcal{S} = \mathcal{M}(\theta_o) \Rightarrow y(t) = \varphi(t)^T\theta_o + e(t),\ e(t) \sim WN(0,\sigma^2)$

If $\mathcal{S} \in \mathfrak{M}$ and $\mathcal{D}_\Theta = \{\theta_o\}$, then $\hat{\theta}_N \sim As\ \mathcal{N}\left(\theta_o, \bar{P}/N\right)$, where:

$$\bar{P} = Var[\varepsilon(t,\theta_o)]\bar{R}^{-1} = \sigma^2\bar{R}^{-1}$$

$$\bar{R} = E\left[\psi(t,\theta_o)\psi(t,\theta_o)^T\right] = E\left[\varphi(t)\varphi(t)^T\right]$$

$$\psi(t,\theta) = -\left[\frac{d}{d\theta}\varepsilon(t,\theta)\right]^T = -\left[-\frac{d}{d\theta}\hat{y}(t,\theta)\right]^T = \varphi(t)$$

since $\hat{y}(t,\theta) = \varphi(t)^T\theta$, $\varepsilon(t,\theta) = y(t) - \hat{y}(t,\theta) = \varphi(t)^T(\theta_o - \theta) + e(t)$

$\bar{P}$ can be directly estimated from N data as: $\bar{P} = \sigma^2\bar{R}^{-1} \approx \hat{P}_N = \hat{\sigma}_N^2\hat{R}_N^{-1}$, with

$$\hat{\sigma}_N^2 = \frac{1}{N}\sum_{t=1}^N \varepsilon(t,\hat{\theta}_N)^2 = \frac{1}{N}\sum_{t=1}^N [y(t) - \varphi(t)^T\hat{\theta}_N]^2$$

$$\hat{R}_N = \frac{1}{N}\sum_{t=1}^N \psi(t,\hat{\theta}_N)\psi(t,\hat{\theta}_N)^T = \frac{1}{N}\sum_{t=1}^N \varphi(t)\varphi(t)^T$$

Note that, under the assumption that $\mathcal{S} \in \mathfrak{M}$, the set $\mathcal{D}_\Theta$ is a singleton that contains the "true" parameter vector θ_o only $\Leftrightarrow$ the matrix $\bar{R} = E\left[\varphi(t)\varphi(t)^T\right]$ is nonsingular

In the case of an $ARX(n_a, n_b)$ model,

$$\varphi(t) = [-y(t-1) \ \cdots \ -y(t-n_a) \ u(t-1) \ \cdots \ u(t-n_b)]^T = \begin{bmatrix} \varphi_y(t) \\ \varphi_u(t) \end{bmatrix}$$

with $\varphi_y(t) = [-y(t-1) \cdots -y(t-n_a)]^T \in \mathbb{R}^{n_a}$, $\varphi_u(t) = [u(t-1) \cdots u(t-n_b)]^T \in \mathbb{R}^{n_b}$

$$\Downarrow$$

$$\bar{R} = E\left[\varphi(t)\varphi(t)^T\right] = E\left[\begin{bmatrix} \varphi_y(t) \\ \varphi_u(t) \end{bmatrix} \begin{bmatrix} \varphi_y(t)^T & \varphi_u(t)^T \end{bmatrix}\right] =$$

$$= E\left[\begin{bmatrix} \varphi_y(t)\varphi_y(t)^T & \varphi_y(t)\varphi_u(t)^T \\ \varphi_u(t)\varphi_y(t)^T & \varphi_u(t)\varphi_u(t)^T \end{bmatrix}\right] = \begin{bmatrix} E\left[\varphi_y(t)\varphi_y(t)^T\right] & E\left[\varphi_y(t)\varphi_u(t)^T\right] \\ E\left[\varphi_u(t)\varphi_y(t)^T\right] & E\left[\varphi_u(t)\varphi_u(t)^T\right] \end{bmatrix}$$

$$= \begin{bmatrix} \bar{R}_{yy}^{(n_a)} & \bar{R}_{yu} \\ \bar{R}_{uy} & \bar{R}_{uu}^{(n_b)} \end{bmatrix} = \begin{bmatrix} \bar{R}_{yy}^{(n_a)} & \bar{R}_{yu} \\ \bar{R}_{yu}^T & \bar{R}_{uu}^{(n_b)} \end{bmatrix}, \text{ where } \bar{R}_{yy}^{(n_a)} = \left[\bar{R}_{yy}^{(n_a)}\right]^T, \ \bar{R}_{uu}^{(n_b)} = \left[\bar{R}_{uu}^{(n_b)}\right]^T$$

For structural reasons, $\bar{R}$ is symmetric and positive semidefinite, since $\forall x \in \mathbb{R}^{n_a+n_b}$:

$$x^T \bar{R} x = x^T E\left[\varphi(t)\varphi(t)^T\right] x = E\left[x^T \varphi(t)\varphi(t)^T x\right] = E\left[\left(x^T \varphi(t)\right)^2\right] \geq 0$$

$$\Downarrow$$

$\bar{R}$ is nonsingular $\Leftrightarrow$ $\bar{R}$ is positive definite (denoted as: $\bar{R} > 0$)

Schur's Lemma: given a symmetric matrix M partitioned as:

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{12}^T & M_{22} \end{bmatrix}$$

(where obviously M_{11} and M_{22} are symmetric), M is positive definite if and only if:

$$M_{22} > 0 \qquad M_{11} - M_{12} M_{22}^{-1} M_{12}^T > 0$$

$$\Downarrow$$

A necessary condition for the invertibility of $\bar{R}$ is that $\bar{R}_{uu} > 0$, i.e., that $\bar{R}_{uu}^{(n_b)}$ is nonsingular, since $\bar{R}_{uu}^{(n_b)}$ is symmetric and positive semidefinite; in fact $\forall x \in \mathbb{R}^{n_b}$:

$$x^T \bar{R}_{uu}^{(n_b)} x = x^T E\left[\varphi_u(t)\varphi_u(t)^T\right] x = E\left[x^T \varphi_u(t)\varphi_u(t)^T x\right] = E\left[\left(x^T \varphi_u(t)\right)^2\right] \geq 0$$

$$\bar{R}_{uu}^{(n_b)} = E\left[\varphi_u(t)\varphi_u(t)^T\right] = E\left[\begin{bmatrix} u(t-1) \\ \vdots \\ u(t-n_b) \end{bmatrix} [u(t-1)\cdots u(t-n_b)]\right] =$$

$$= \begin{bmatrix} E\left[u(t-1)^2\right] & E\left[u(t-1)u(t-2)\right] & \cdots & E\left[u(t-1)u(t-n_b)\right] \\ E\left[u(t-2)u(t-1)\right] & E\left[u(t-2)^2\right] & \cdots & E\left[u(t-2)u(t-n_b)\right] \\ \vdots & \vdots & \ddots & \vdots \\ E\left[u(t-n_b)u(t-1)\right] & E\left[u(t-n_b)u(t-2)\right] & \cdots & E\left[u(t-n_b)^2\right] \end{bmatrix} =$$

$$= \begin{bmatrix} r_u(t-1,0) & r_u(t-1,1) & \cdots & r_u(t-1,n_b-1) \\ r_u(t-1,1) & r_u(t-2,0) & \cdots & r_u(t-2,n_b-2) \\ \vdots & \vdots & \ddots & \vdots \\ r_u(t-1,n_b-1) & r_u(t-2,n_b-2) & \cdots & r_u(t-n_b,0) \end{bmatrix} =$$

$$= \begin{bmatrix} r_u(0) & r_u(1) & \cdots & r_u(n_b-1) \\ r_u(1) & r_u(0) & \cdots & r_u(n_b-2) \\ \vdots & \vdots & \ddots & \vdots \\ r_u(n_b-1) & r_u(n_b-2) & \cdots & r_u(0) \end{bmatrix}$$

where $r_u(t,\tau) = E[u(t)u(t-\tau)]$ is the correlation function of the input $u(\cdot)$, which is independent of t for any stationary process $u(\cdot)$: $r_u(t_1,\tau) = r_u(t_2,\tau) = r_u(\tau), \ \forall t_1,t_2,\tau$

A stationary signal $u(\cdot)$ is **persistently exciting of order $\mathbf{n}$** $\Leftrightarrow \bar{R}_{uu}^{(n)}$ is nonsingular

Examples:

- the discrete-time unitary impulse $u(t) = \delta(t) = \begin{cases} 1, & \text{if } t = 1 \\ 0, & \text{if } t \neq 1 \end{cases}$

  is not persistently exciting of any order, since $r_u(\tau) = 0, \ \forall \tau \Rightarrow \bar{R}_{uu}^{(n)} = 0_{n \times n}$

- the discrete-time unitary step $u(t) = \varepsilon(t) = \begin{cases} 1, & \text{if } t = 1, 2, \ldots \\ 0, & \text{if } t = \ldots, -1, 0 \end{cases}$

  is persistently exciting of order 1 only, since $r_u(\tau) = 1, \ \forall \tau \Rightarrow \bar{R}_{uu}^{(n)} = 1_{n \times n}$

- the discrete-time signal $u(t)$ consisting of m different sinusoids:

  $$u(t) = \sum_{k=1}^{m} \mu_k \cos(\omega_k t + \varphi_k), \quad \text{where } 0 \leq \omega_1 < \omega_2 < \ldots < \omega_m \leq \pi$$

  is persistently exciting of order $n = \begin{cases} 2m, & \text{if } 0 < \omega_1 \text{ and } \omega_m < \pi \\ 2m-1, & \text{if } 0 = \omega_1 \text{ or } \omega_m = \pi \\ 2m-2, & \text{if } 0 = \omega_1 \text{ and } \omega_m = \pi \end{cases}$

- the white noise $u(t) \sim WN(0, \sigma^2)$ is persistently exciting of all orders, since $r_u(0) = \sigma^2$ and $r_u(\tau) = 0, \ \forall \tau \neq 0 \Rightarrow \bar{R}_{uu}^{(n)} = \sigma^2 I_n$

As a consequence, a necessary condition for the invertibility of $\bar{R}$ is that the signal $u(\cdot)$ is persistently exciting of order n_b at least

$$\Downarrow$$

A necessary condition to univocally estimate the parameters of an $ARX(n_a, n_b)$ (i.e., to prevent any problem of *experimental identifiability* related to the choice of u) is that the signal $u(\cdot)$ is persistently exciting of order n_b at least

The matrix $\bar{R}$ may however be singular also for problems of *structural identifiability* related to the choice of the model class $\mathfrak{M}$: in fact, even in the case $\mathcal{S} \in \mathfrak{M}$, if $\mathfrak{M}$ is redundant or *overparametrized* (i.e., its orders are greater than necessary), then an infinite number of models may represent $\mathcal{S}$ by means of suitable pole-zero cancellations in the denominator and numerator of the involved transfer functions

$$\Downarrow$$

To summarize, only in the case that $\mathcal{S} = \mathcal{M}(\theta_o)$ is an $ARX(n_a, n_b)$ (without any pole-zero cancellation in the transfer function) and $\mathfrak{M}$ is the class of $ARX(n_a, n_b)$ models, if the input signal $u(\cdot)$ is persistently exciting of order n_b at least, then the least-squares estimate $\hat{\theta}_N$ asymptotically converges to the "true" parameter vector θ_o

Least-squares method: practical procedure

1. Starting from N data points of $u(\cdot)$ and $y(\cdot)$, build the regression vector $\varphi(t)$ and the matrix $\hat{R}_N = \frac{1}{N}\sum_{t=1}^{N}\varphi(t)\varphi(t)^T \xrightarrow[N\to\infty]{} \bar{R}$ if $\varphi(\cdot)$ is stationary;

   in compact matrix form, $\hat{R}_N = \frac{1}{N}\Phi^T\Phi$, where $\Phi = \begin{bmatrix} \varphi(1)^T \\ \vdots \\ \varphi(N)^T \end{bmatrix}$

2. Check if $\hat{R}_N$ is nonsingular, i.e., if $\det \hat{R}_N \neq 0$: if there exists the matrix $\hat{R}_N^{-1}$, then the estimate is unique and it is given by: $\hat{\theta}_N = \hat{R}_N^{-1}\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\,y(t)$;

   in a matrix form, $\hat{\theta}_N = \hat{R}_N^{-1}\frac{1}{N}\Phi^T y = \left(\Phi^T\Phi\right)^{-1}\Phi^T y$, where $y = \begin{bmatrix} y(1) \\ \vdots \\ y(N) \end{bmatrix}$

3. Evaluate the prediction error of the estimated model $\varepsilon(t,\hat{\theta}_N) = y(t) - \varphi(t)^T\hat{\theta}_N$ and approximate the estimate uncertainty as: $\Sigma_{\hat{\theta}_N} = \hat{R}_N^{-1}\frac{1}{N^2}\sum_{t=1}^{N}\varepsilon(t,\hat{\theta}_N)^2$, where the elements on the diagonal are the variances of each parameter $[\hat{\theta}_N]_i$

4. Check the whiteness of $\varepsilon(t,\hat{\theta}_N)$ by means of a suitable test

Anderson's whiteness test

Let $\varepsilon(\cdot)$ be the signal under test and N be the (sufficiently large) number of samples

1. Compute the sample correlation function $\hat{r}_\varepsilon(\tau) = \frac{1}{N}\sum_{t=\tau+1}^{N} \varepsilon(t)\varepsilon(t-\tau),\ 0 \le \tau \le \bar{\tau}$ $(\bar{\tau} = 25 \text{ or } 30)$, and the normalized sample correlation function $\hat{\rho}_\varepsilon(\tau) = \frac{\hat{r}_\varepsilon(\tau)}{\hat{r}_\varepsilon(0)} \Rightarrow$ if $\varepsilon(\cdot)$ is white with zero mean, then $\hat{\rho}_\varepsilon(\tau)$ is asymptotically normally distributed:
$$\hat{\rho}_\varepsilon(\tau) \sim As\ \mathcal{N}\left(0, \tfrac{1}{N}\right), \quad \forall \tau > 0$$
moreover, $\hat{\rho}_\varepsilon(\tau_1)$ and $\hat{\rho}_\varepsilon(\tau_2)$ are asymptotically uncorrelated $\forall \tau_1 \neq \tau_2$
2. Fix a confidence level α, i.e., the probability α that asymptotically $|\hat{\rho}_\varepsilon(\tau)| \le \beta$, and evaluate β; in particular, it turns out that: $\beta = \begin{cases} 1/\sqrt{N}, & \text{for } \alpha = 68.3\% \\ 2/\sqrt{N}, & \text{for } \alpha = 95.4\% \\ 3/\sqrt{N}, & \text{for } \alpha = 99.7\% \end{cases}$
3. The test is failed if the number of τ values such that $|\hat{\rho}_\varepsilon(\tau)| \le \beta$ is less than $\lfloor \alpha\bar{\tau} \rfloor$, where $\lfloor x \rfloor$ denotes the biggest integer less than or equal to x, otherwise it is passed

Recursive least-squares methods

The least-squares estimate referred to a generic time instant t is given by:

$$\hat{\theta}_t = \left[\sum_{i=1}^{t} \varphi(i)\,\varphi(i)^T\right]^{-1} \sum_{i=1}^{t} \varphi(i)\,y(i) = S(t)^{-1} \sum_{i=1}^{t} \varphi(i)\,y(i)$$

where

$$S(t) = \sum_{i=1}^{t} \varphi(i)\varphi(i)^T = \sum_{i=1}^{t-1} \varphi(i)\varphi(i)^T + \varphi(t)\varphi(t)^T = S(t-1) + \varphi(t)\varphi(t)^T$$

The least-squares estimate referred to the time instant $t-1$ is given by:

$$\hat{\theta}_{t-1} = \left[\sum_{i=1}^{t-1} \varphi(i)\,\varphi(i)^T\right]^{-1} \sum_{i=1}^{t-1} \varphi(i)\,y(i) = S(t-1)^{-1} \sum_{i=1}^{t-1} \varphi(i)\,y(i)$$

and then:

$$\begin{aligned}
\hat{\theta}_t &= S(t)^{-1} \sum_{i=1}^{t} \varphi(i)\,y(i) = S(t)^{-1} \left[\sum_{i=1}^{t-1} \varphi(i)\,y(i) + \varphi(t)\,y(t)\right] = \\
&= S(t)^{-1}[S(t-1)\hat{\theta}_{t-1} + \varphi(t)\,y(t)] = \\
&= S(t)^{-1}\{[S(t) - \varphi(t)\,\varphi(t)^T]\hat{\theta}_{t-1} + \varphi(t)\,y(t)\} = \\
&= \hat{\theta}_{t-1} - S(t)^{-1}\varphi(t)\,\varphi(t)^T\,\hat{\theta}_{t-1} + S(t)^{-1}\varphi(t)\,y(t) = \\
&= \hat{\theta}_{t-1} + S(t)^{-1}\varphi(t)\,[y(t) - \varphi(t)^T\,\hat{\theta}_{t-1}]
\end{aligned}$$

Since the estimate can be computed as: $\hat{\theta}_t = \hat{\theta}_{t-1} + S(t)^{-1}\varphi(t)[y(t) - \varphi(t)^T\hat{\theta}_{t-1}]$, a first *recursive least-squares (RLS)* algorithm (denoted as *RLS-1*) is the following one:

$$\begin{aligned} S(t) &= S(t-1) + \varphi(t)\,\varphi(t)^T && \text{(time update)} \\ K(t) &= S(t)^{-1}\varphi(t) && \text{(algorithm gain)} \\ \varepsilon(t) &= y(t) - \varphi(t)^T\,\hat{\theta}_{t-1} && \text{(prediction error)} \\ \hat{\theta}_t &= \hat{\theta}_{t-1} + K(t)\varepsilon(t) && \text{(estimate update)} \end{aligned}$$

An alternative algorithm is derived by considering the matrix $R(t) = \frac{1}{t}\sum_{i=1}^{t}\varphi(i)\varphi(i)^T$:

$$\begin{aligned} R(t) &= \tfrac{1}{t}S(t) = \tfrac{1}{t}S(t-1) + \tfrac{1}{t}\varphi(t)\,\varphi(t)^T = \\ &= \left(\tfrac{1}{t} + \tfrac{1}{t-1} - \tfrac{1}{t-1}\right) S(t-1) + \tfrac{1}{t}\varphi(t)\,\varphi(t)^T = \\ &= \tfrac{1}{t-1}S(t-1) + \left(\tfrac{1}{t} - \tfrac{1}{t-1}\right) S(t-1) + \tfrac{1}{t}\varphi(t)\,\varphi(t)^T = \\ &= R(t-1) + \tfrac{t-1-t}{t(t-1)}S(t-1) + \tfrac{1}{t}\varphi(t)\,\varphi(t)^T = \\ &= R(t-1) - \tfrac{1}{t}R(t-1) + \tfrac{1}{t}\varphi(t)\,\varphi(t)^T = \\ &= \left(1 - \tfrac{1}{t}\right) R(t-1) + \tfrac{1}{t}\varphi(t)\,\varphi(t)^T \end{aligned}$$

A second recursive least-squares algorithm (denoted as *RLS-2*) is then the following one:

$$R(t) = \left(1 - \tfrac{1}{t}\right) R(t-1) + \tfrac{1}{t}\varphi(t)\,\varphi(t)^T \qquad \text{(time update)}$$
$$K(t) = \tfrac{1}{t} R(t)^{-1}\varphi(t) \qquad \text{(algorithm gain)}$$
$$\varepsilon(t) = y(t) - \varphi(t)^T\,\hat{\theta}_{t-1} \qquad \text{(prediction error)}$$
$$\hat{\theta}_t = \hat{\theta}_{t-1} + K(t)\varepsilon(t) \qquad \text{(estimate update)}$$

The main drawback of *RLS-1* and *RLS-2* algorithms is the inversion at each step of the square matrices $S(t)$ and $R(t)$, respectively, whose dimensions are equal to the number of estimated parameters $\Rightarrow$ by applying the Matrix Inversion Lemma:

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$$

taking $A = S(t-1)$, $B = D^T = \varphi(t)$, $C = 1$ and introducing $V(t) = S(t)^{-1}$ gives:

$$V(t) = S(t)^{-1} = \left[S(t-1) + \varphi(t)\,\varphi(t)^T\right]^{-1} =$$
$$= S(t-1)^{-1} - S(t-1)^{-1}\varphi(t)\Big[1 + \underbrace{\varphi(t)^T S(t-1)^{-1}\varphi(t)}_{\text{it is a scalar}}\Big]^{-1}\varphi(t)^T S(t-1)^{-1} =$$
$$= V(t-1) - \left[1 + \varphi(t)^T V(t-1)\varphi(t)\right]^{-1} V(t-1)\varphi(t)\,\varphi(t)^T V(t-1)$$

Since $V(t) = S(t)^{-1} = V(t-1) - \left[1 + \varphi(t)^T V(t-1) \varphi(t)\right]^{-1} V(t-1) \varphi(t) \varphi(t)^T V(t-1)$ a third recursive least-squares algorithm (denoted as *RLS-3*) is then the following one:

$$\begin{aligned}
\beta_{t-1} &= 1 + \varphi(t)^T V(t-1) \varphi(t) && \text{(scalar weight)} \\
V(t) &= V(t-1) - \beta_{t-1}^{-1} V(t-1) \varphi(t) \varphi(t)^T V(t-1) && \text{(time update)} \\
K(t) &= V(t) \varphi(t) && \text{(algorithm gain)} \\
\varepsilon(t) &= y(t) - \varphi(t)^T \hat{\theta}_{t-1} && \text{(prediction error)} \\
\hat{\theta}_t &= \hat{\theta}_{t-1} + K(t) \varepsilon(t) && \text{(estimate update)}
\end{aligned}$$

To use the recursive algorithms, initial values for their start-up are obviously required; in the case of the *RLS-3* algorithm:

- the correct initial conditions, at a time instant t_o when $S(t_o)$ becomes invertible, are:

$$V(t_o) = S(t_o)^{-1} = \left[\sum_{i=1}^{t_o} \varphi(i)\, \varphi(i)^T \right]^{-1}, \quad \hat{\theta}_{t_o} = V(t_o) \sum_{i=1}^{t_o} \varphi(i)\, y(i)$$

- assuming $n = \dim(\theta)$, a much simpler alternative is to use:

$$V(0) = \alpha I_n, \ \alpha > 0, \quad \text{and} \quad \hat{\theta}_0 = 0_{n \times 1}$$

$\hat{\theta}_t$ rapidly changes from $\hat{\theta}_0$ if $\alpha \approx 1$, while $\hat{\theta}_t$ slowly changes from $\hat{\theta}_0$ if $\alpha \ll 1$

Model structure selection and validation

A most natural approach to search for a suitable model structure $\mathfrak{M}$ is simply to test a number of different ones and to compare the resulting models

Given a model $\mathcal{M}(\theta) \in \mathfrak{M}$ with complexity $n = \dim(\theta)$, the cost function

$$J(\theta)^{(n)} = \frac{1}{N}\sum_{t=1}^{N} \varepsilon(t)^2 = \frac{1}{N}\sum_{t=1}^{N} \left(y(t) - \hat{y}(t,\theta)\right)^2$$

provides a measure of the fitting of the data set y provided by $\mathcal{M}(\theta) \Rightarrow$

if $\hat{\theta}_N = \operatorname{argmin} J(\theta)^{(n)}$, then $J(\hat{\theta}_N)^{(n)}$ measures the best fitting of data y provided by $\mathfrak{M}$ and represents a subjective (and very optimistic) evaluation of the quality of $\mathfrak{M}$

In order to perform a more objective evaluation, it would be necessary to measure the model class accuracy on data different from those used in the identification $\Rightarrow$ to this purpose, there are different criteria:

- Cross-Validation
- Akaike's Final Prediction-Error Criterion (FPE)
- Model Structure Selection Criteria: AIC and MDL (or BIC)

Cross-Validation

If the overall data set is sufficiently huge, it can be partitioned into two subsets:

- the *estimation data* are the ones used to estimate the model $\mathcal{M}(\hat{\theta}_N) \in \mathfrak{M}$
- the *validation data* are the ones that have not been used to build any of the models we would like to compare

For any given model class $\mathfrak{M}$, first the model $\mathcal{M}(\hat{\theta}_N)$ that better reproduces the estimation data is identified, and then its performance is evaluated by computing the mean square error on the validation data only: the model that minimizes such a criterion among different classes $\mathfrak{M}$ is chosen as the most suitable one

It can be noted that, within any model class, higher order models usually suffer from *overfitting*, i.e., they fit too much the estimation data to fit also the noise term and then their predictive capability on a fresh data set (corrupted by a different noise) is smaller with respect to lower order models

Akaike's Final Prediction-Error Criterion (FPE)

In order to consider any possible realization of data $y(t,s)$ that depends on the outcome s of the random experiment, let us consider as objective criterion:

$$\bar{J}(\theta) = E[(y(t,s) - \hat{y}(t,s,\theta))^2]$$

Since $\hat{\theta}_N = \hat{\theta}_N(s)$ depends on a particular data set $y(t,s)$ generated by a particular outcome s, the *Final Prediction Error (FPE)* criterion is defined as the mean on any possible outcome s:

$$FPE = E[\bar{J}(\hat{\theta}_N(s))]$$

In the case of the *AR* model class, it can be proved that:

$$FPE = \frac{N+n}{N-n} J(\hat{\theta}_N)^{(n)}$$

where $J(\hat{\theta}_N)^{(n)}$ is a monotonic decreasing function of n while $\frac{N+n}{N-n} \to \infty$ as $n \to N$

$\Rightarrow FPE$ is decreasing for lower values of n and it is increasing for higher values of n

$\Rightarrow$ the optimal model complexity corresponds to the minimum of FPE

The same formula is usually used also in the case of other model classes (*ARX*, *ARMAX*)

Akaike's Information Criterion (AIC)

Such a criterion is derived on the basis of statistical considerations and aims at minimizing the so-called Kullback distance between the "true" probability density function of the data and the p.d.f. produced by a given model $\mathcal{M}(\hat{\theta}_N)$:

$$AIC = n\frac{2}{N} + \ln J(\hat{\theta}_N)^{(n)}$$

The optimum model order n^* minimizes the *AIC* criterion: $n^* = \arg\min AIC$

For large values of N, the *FPE* and *AIC* criteria lead to the same result:

$$\begin{aligned}\ln FPE &= \ln \tfrac{N+n}{N-n} J(\hat{\theta}_N)^{(n)} = \ln \tfrac{1+n/N}{1-n/N} J(\hat{\theta}_N)^{(n)} = \\ &= \ln(1+n/N) - \ln(1-n/N) + \ln J(\hat{\theta}_N)^{(n)} \cong \\ &\cong n/N - (-n/N) + \ln J(\hat{\theta}_N)^{(n)} = n\tfrac{2}{N} + \ln J(\hat{\theta}_N)^{(n)} = AIC\end{aligned}$$

AIC criterion is directed to find system descriptions that give the smallest mean-square error: a model that apparently gives a smaller mean-square (prediction) error fit will be chosen even if it is quite complex

Rissanen's Minimum Description Length Criterion (MDL)

In practice, one may want to add an extra penalty for the model complexity, in order to reflect the cost of using it

What is meant by a complex model and what penalty should be associated with are usually subjective issues; an approach that is conceptually related to code theory and information measures has been taken by Rissanen, who stated that a model should be sought that allows the shortest possible code or description of the observed data, leading to the *Minimum Description Length (MDL)* criterion:

$$MDL = n\frac{\ln N}{N} + \ln J(\hat{\theta}_N)^{(n)}$$

As in the *AIC* criterion, the model complexity penalty is proportional to n; however, while in *AIC* the constant is $\frac{2}{N}$, in *MDL* the constant is $\frac{\ln N}{N} > \frac{2}{N}$ for any $N \geq 8$ $\Rightarrow$ the *MDL* criterion leads to much more parsimonious models than those selected by the *AIC* criterion, especially for large values of N

Such a criterion has also been termed *BIC* by Akaike