MIDA : MODEL IDENTIFICATION & DATA ANALYSIS

02/03/2019

24/02 - 6/04 BITTANTI

15/04 - 5/06 SAVARESI

Exercices sessions : RAPIZZA

Program :

1. Prediction with input/output models
2. Predictive identification

] Kolmogorov - Wiener input/output models

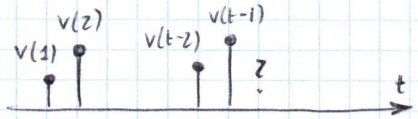
3. Kalman filtering and prediction
4. Non-parametric identification

] Kalman state-space models (not I/O)

PART I : PREDICTION

INTRODUCTION

(Motivating example). Consider a given sequence of real numbers ordered by time t (discrete time): $v(t-1), v(t-2), \dots, v(1)$.
Can we estimate $v(t)$ knowing $\{v(t-1), v(t-2), \dots, v(1)\}$?



(suppose we don't know how the data have been generated).

We have to invent a predictor $\hat{v}(t|t-1)$

estimate of $v(t)$ (unknown) given the past values of v up to $t-1$: $v(t-1), v(t-2), \dots, v(1)$

it can be any combination of the past data

We also have to invent a law of prediction ($\equiv$ form of the predictor).
For instance we can impose:

$$\hat{v}(t|t-1) = a_1 v(t-1) + a_2 v(t-2) + \dots + a_n v(t-n)$$

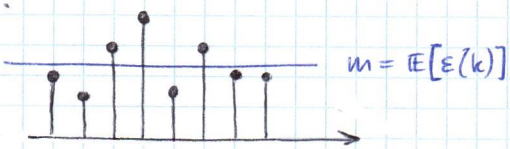
Problem: Find $a_1, \dots, a_n$ (the parameters) and decide n (for the moment is given).
How to set $a_1, \dots, a_n$?
We consider the basic criterion: we consider the **PREDICTION ERROR** at time k:

$$\underbrace{\varepsilon(k)}_{\text{prediction error}} = \underbrace{v(k)}_{\text{real value}} - \underbrace{\hat{v}(k|k-1)}_{\text{prediction}}$$

The hope is to have $\varepsilon(k) = 0 \ \forall k$ but it's utopian.

- Case 1:

$\varepsilon(k)$:

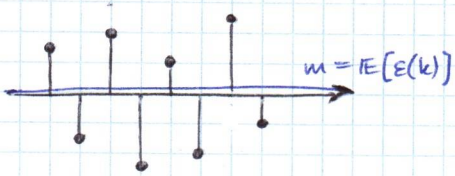


It's not a good prediction error since $m \neq 0$. There is a sistematic error.
We want the mean of the prediction error to be zero.
A better predictor than this (called $\hat{v}(k|k-1)$) can be:

$$\begin{cases} \hat{v}(k|k-1) = \hat{v}(k|k-1) - m \\ \varepsilon(k) = v(k) - \hat{v}(k|k-1) \end{cases}$$

- Case 2:

$\varepsilon(k)$:



This time is good for the mean, but there is another problem: the sign always change from one step to another, so there is a regularity in the prediction error ($\Rightarrow$ it's not a good predictor). A better predictor than this would be:

$$\hat{v}(k|k-1) > \hat{v}(k|k-1) \quad \text{if } \varepsilon(\text{previous point is negative})$$

KEY IDEA: We can always find a better predictor whenever the predictor error has a regularity. When the predictor will be optimal? When the corresponding prediction error does not exhibit any regularity, does not contain any hidden dynamic. So, a predictor is good if the corresponding prediction error is a zero-mean unpredictable signal (all the regularities are in the model, the model captured all the dynamics and we can't do better).

UNPREDICTABLE SIGNAL ≡ WHITE NOISE : an unpredictable signal is a signal such that the past has no role in predicting the future. How can we characterize it? (We'll come back to that).

Suppose that $\hat{v}$ is s.t. the corresponding prediction error $\varepsilon(k)$ is white noise:

$$v(t) = \underbrace{\hat{v}(t|t-1)}_{\text{prediction}} + \underbrace{\varepsilon(t)}_{\text{prediction error}}$$

$$\Longrightarrow \quad v(t) = a_1 v(t-1) + a_2 v(t-2) + \dots + a_n v(t-n) + \varepsilon(t)$$

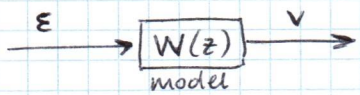
We now introduce the operator z: $\quad z^k v(t) = v(t+k) \qquad k \in \mathbb{Z}$

$$\Longrightarrow \quad v(t) = a_1 z^{-1} v(t) + a_2 \quad v(t) + \dots + a_n z^{-n} v(t) + \varepsilon(t)$$

$$\Longrightarrow \quad \underbrace{(1 - a_1 z^{-1} - a_2 z^{-2} - \dots - a_n z^{-n})}_{:= A(z)} v(t) = \varepsilon(t)$$

$$\Longrightarrow \quad v(t) = \frac{1}{A(z)} \varepsilon(t) := \overset{*}{W(z)} \varepsilon(t) \quad :$$

:= TRANSFER FUNCTION from ε to v



So the PREDICTION PROBLEM corresponds to find a good predictor that describe the signal v as an output of a dynamical system fed by white noise (namely a fully unpredictable signal).

$$* \quad W(z) = \frac{1}{1 - a_1 z^{-1} - \dots - a_n z^{-n}} = \frac{z^n}{z^n - a_1 z^{n-1} - \dots - a_n}$$

We call:

- ZEROS of the transfer function the complex numbers z s.t. the numerator is $= 0$
- POLES of the transfer function the complex numbers z s.t. the denominator is $= 0$

OBJECTIVE: Characterization of the notion of UNPREDICTABLE SIGNAL (≡ WHITE NOISE)

Next points:

- Random variable
- Random vector
- Random process (/ Stochastic process)
- Unpredictable signal (/ White noise)

Refresh of **RANDOM VARIABLE**

A Random variable is a real variable depending upon the outcome of a random experiment.

Example: tossing a coin. We define the random variable v as:

$$v = \begin{cases} +1 & \text{head} \\ -1 & \text{tails} \end{cases}$$

(real and depending on the random process of tossing a coin)
If s denotes the outcome of the experiment, we can write the random variable as $v(s)$, where v is a real number and s is the outcome of the experiment.

Characterization of a random variable:

- mean ($\equiv$ expected value): $\mathbb{E}[v] = m$
- variance: $Var(v) = \mathbb{E}[(v - \mathbb{E}[v])^2] = \lambda^2 = \sigma^2$
- standard deviation: $\sqrt{Var(v)} = \lambda = \sigma$

Example: $v \sim G(m, \lambda^2)$ $\quad (\sim N(m, \lambda^2))$

$\Longrightarrow$ $v(s)$ belongs to $(m - 2\lambda, m + 2\lambda)$ in 95%

Refresh of RANDOM VECTOR

A random vector is a vector whose elements are random variables.
For instance, if we have two random variables v_1, v_2 we can define the vector: $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$

Characterization of random vector:

- mean value ($\equiv$ expected value):

$$\left.\begin{aligned} \mathbb{E}[v_1] &= m_1 \\ \mathbb{E}[v_2] &= m_2 \end{aligned}\right\} \Longrightarrow \mathbb{E}[v] = \begin{bmatrix} \mathbb{E}[v_1] \\ \mathbb{E}[v_2] \end{bmatrix}, \quad m = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}$$

- variance: $\{Var(v_1), Var(v_2)\}$ is not enough, we want to know the cross variance between the random variables:

Cross-variance $v_1 - v_2$: $\Delta = \mathbb{E}[(v_1 - \mathbb{E}[v_1])(v_2 - \mathbb{E}[v_2])]$

Cross-variance $v_2 - v_1$: $\nabla = \mathbb{E}[(v_2 - \mathbb{E}[v_2])(v_1 - \mathbb{E}[v_1])]$

$\Longrightarrow \nabla = \Delta$

We have a main definition:

(CO) VARIANCE MATRIX $:= Var(v) = \begin{bmatrix} Var(v_1) & \Delta \\ \nabla & Var(v_2) \end{bmatrix}$

Note that $Var(v)$ is symmetric since $\nabla = \Delta$.

Alternatively we can write: $Var(v) = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix}$

where: $\begin{cases} \lambda_{11} = Var(v_1) & \lambda_{12} = \Delta \\ \lambda_{22} = Var(v_2) & \lambda_{21} = \nabla \end{cases}$

Observation:

We define a new random vector $\Delta v := \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} - \begin{bmatrix} \mathbb{E}[v_1] \\ \mathbb{E}[v_2] \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} - \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}$

Prop. $Var(v) = \mathbb{E}[(\Delta v) \cdot (\Delta v)^T] := \mathbb{E}[(\Delta v) \cdot (\Delta v)']$

proof.

$$\Delta v = \begin{bmatrix} v_1 - m_1 \\ v_2 - m_2 \end{bmatrix} \Longrightarrow \Delta v \cdot (\Delta v)' = \begin{bmatrix} v_1 - m_1 \\ v_2 - m_2 \end{bmatrix} \begin{bmatrix} v_1 - m_1 & v_2 - m_2 \end{bmatrix}$$

$$= \begin{bmatrix} (v_1 - \mathbb{E}[v_1])^2 & (v_1 - \mathbb{E}[v_1])(v_2 - \mathbb{E}[v_2]) \\ (v_2 - \mathbb{E}[v_2])(v_1 - \mathbb{E}[v_1]) & (v_2 - \mathbb{E}[v_2])^2 \end{bmatrix}$$

$$\Longrightarrow Var(v) = \mathbb{E}[\Delta v \cdot (\Delta v)']$$

■

Important property of variance:

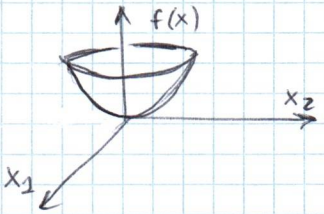
- scalar case (random variable): $Var(v) = \mathbb{E}[(v - \mathbb{E}[v])^2] \geq 0$
- vector case (random vector): $Var(v) \geq 0$

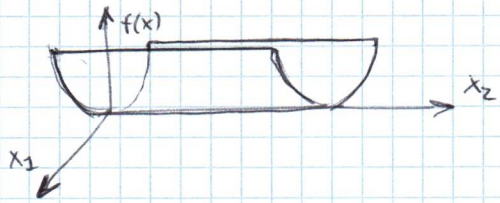
  the matrix is POSITIVE SEMI-DEFINITE

Refresh on POSITIVE SEMI-DEFINITE MATRIX:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \in \mathbb{R}^{2\times 2}, \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2, \quad x' = [x_1 \ \ x_2]$$

$$\Longrightarrow \quad f(x) := x'Ax = [x_1 \ \ x_2] \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$





- A is positive semi-definite when $f(x) \geq 0 \quad \forall x$
- A is positive definite when $f(x) \geq 0 \quad \forall x$ and $f(x) = 0$ only for $x = 0$

Prop. Any variance matrix is positive semi-definite

proof.

$\forall A$ variance matrix can be written as: $A = \mathbb{E}[\Delta v \cdot (\Delta v)']$

$$\Longrightarrow \quad x'Ax = x' \, \mathbb{E}[\Delta v \cdot (\Delta v)'] \, x = \mathbb{E}[x'(\Delta v \cdot (\Delta v)')x]$$
$$= \mathbb{E}[(x'\Delta v)^2] \geq 0 \qquad \blacksquare$$

As a consequence of the property, since: A positive semi-def. $\Rightarrow \det(A) \geq 0$:

$$Var(v) = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix} \Longrightarrow \det(Var(v)) = \lambda_{11}\lambda_{22} - \lambda_{12}^2 \geq 0 \qquad (*)$$

We define the COVARIANCE COEFFICIENT (/Correlation coefficient) as:

$$\rho = \frac{\lambda_{12}}{\sqrt{\lambda_{11}}\sqrt{\lambda_{22}}}$$

From $(*)$ we have: $|\rho| \leq 1 \longrightarrow \rho \in [-1, +1]$

Example: Consider two random variables v_1, v_2:

$$v_1 \sim (0, \lambda_{11}), \quad v_2 = \alpha v_1, \quad \alpha \in \mathbb{R} \setminus \{0\}$$

What is the covariance coefficient between v_1 and v_2?

- Compute $\mathbb{E}[v_2]$ $\quad (\mathbb{E}[\alpha_1 v_1 + \alpha_2 v_2] = \alpha_1 \mathbb{E}[v_1] + \alpha_2 \mathbb{E}[v_2] \quad \alpha_i \in \mathbb{R}, \ v_i \text{ R.V.})$

  $$\mathbb{E}[v_2] = \mathbb{E}[\alpha v_1] = \alpha \, \mathbb{E}[v_1] = 0$$

- Compute cross variance $v_1 - v_2$:

  $$\lambda_{12} = \mathbb{E}[v_1 v_2] = \mathbb{E}[v_1(\alpha v_1)] = \alpha \, \mathbb{E}[v_1^2] = \alpha \lambda_{11}$$

  $$\lambda_{22} = \mathbb{E}[(v_2 - \cancel{\mathbb{E}[v_2]})^2] = \mathbb{E}[v_2^2] = \mathbb{E}[\alpha^2 v_1^2]$$
  $$= \alpha^2 \, \mathbb{E}[v_1^2] = \alpha^2 \lambda_{11}^2$$

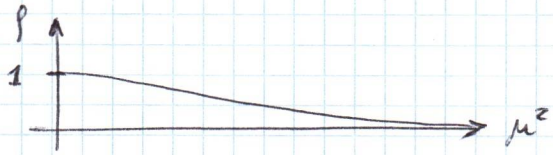
$$\rho = \frac{\lambda_{12}}{\sqrt{\lambda_{11}}\sqrt{\lambda_{22}}} = \frac{\alpha\lambda_{11}}{\sqrt{\lambda_{11}}\sqrt{\alpha^2\lambda_{11}^2}} = \begin{cases} +1 & \alpha > 0 \\ -1 & \alpha < 0 \end{cases}$$

Intuitive explaination: $v_2 = \alpha v_1 \Rightarrow$ if one knows the value taken by the random variable v_1 then one can say what is the value of v_2 without any error ($|\rho| = 1$).
Viceversa, if $|\rho| = 0$, knowing one value (the value of one variable) is useless to predict the value of the second variable $\Rightarrow$ if $\rho = 0$ we can say that v_1 and v_2 are uncorrelated.

Example: Consider $v_2 = \alpha v_1 + \eta$, $\eta \sim (0, \mu^2)$ noise.

if $\mu^2 = 0 \Rightarrow v_2 = \alpha v_1$

If $\mu^2 \neq 0 \Rightarrow$ there is noise, if it's large then it's difficult to estimate v_2 from v_1:



$$\rho \xrightarrow{\mu^2 \to \infty} 0$$

Refresh on STOCHASTIC PROCESS (≡ RANDOM PROCESS)

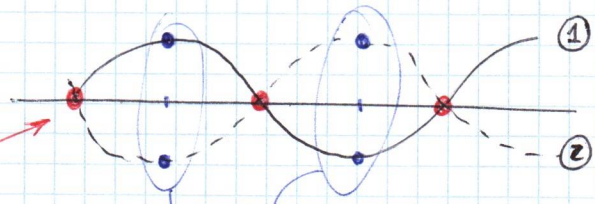
A stochastic/random process is an infinite sequence of random variables:

$v(t)$, t discrete time from $-\infty$ to $+\infty$.

Note that it's not a function of t, is a set of functions because the random variable depends on the outcome of a random experiment

$\Longrightarrow$ $v(t,s)$ (t discrete time from $-\infty$ to $+\infty$
s outcome of a random experiment)

For example: $v(t) = \begin{cases} \sin(t) & \text{if head} \\ -\sin(t) & \text{if tails} \end{cases}$:



there are two possible functions, which are called REALIZATIONS

Var = 0 in these points

here the variance is not 0 since in the case 1 we have one value and in the case 2 we have another value

Characterization of a stochastic process:

- mean value: $m(t) = \mathbb{E}[v(t,s)]$
- variance: $\lambda^2(t) = Var(v(t)) = \mathbb{E}[(v(t) - m(t))^2]$
- covariance function:

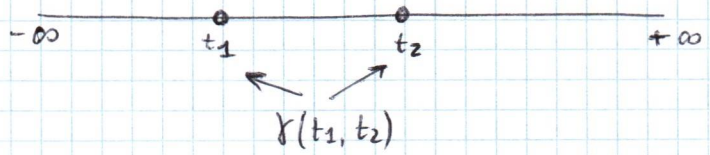
$$\gamma(t_1, t_2) = \mathbb{E}[(v(t_1) - m(t_1))(v(t_2) - m(t_2))]$$

cross variance between random variables of the process at time t_1 and t_2

Properties of the covariance function:

- $t_2 = t_1 = t \Rightarrow \gamma(t_1, t_2) = \gamma(t,t) = \mathbb{E}[(v(t) - m(t))^2] = Var(v(t))$ variance of the process
- $\gamma(t_1, t_2) = \gamma(t_2, t_1)$

The covariance function represents the cross variance between the random variables $v(t_1)$ and $v(t_2)$:



FOCUS ON A CLASS OF RANDOM PROCESSES: **STATIONARY PROCESSES**

A stationary process is a stochastich process s.t. :

- $\mathbb{E}[v(t,s)] = m(t) = \text{constant} = m$
- $\lambda^2(t) = \text{constant} = \lambda^2$
- $\gamma(t_1, t_2)$ function only of $t_2 - t_1 \implies \gamma(t_1, t_2) = \gamma(t_2 - t_1)$
  (the correlations of these processes depend only on how far are the two time points t_1 and t_2)

$\tau := t_2 - t_1 \implies \gamma(t_1, t_2) = \gamma(\tau)$

Observations :

- $\gamma(0) = \gamma(t,t) = \mathbb{E}[(v(t) - \mathbb{E}[v(t)])^2] = Var(v(t)) = \lambda^2(t) = \lambda^2$
  variance of the process
  (Note that $\gamma(t,t)$ is the variance of the process at time t, so, since $\gamma(0) = \gamma(t,t)$ is constant $\implies$ in a stationary random process $\gamma(0)$ is the process variance and it does not depend on time)
- $|\gamma(\tau)| \leq \gamma(0)$
  the covariance function cannot exceed (in absolute value) the variance of the process

  proof.

  We consider t_1, t_2 and we define two random variables :

  $$\begin{cases} v_1 = v(t_1) \\ v_2 = v(t_2) \end{cases}$$

  We know that the covariance coefficient of two random variables is always less than 1 in absolute value:

  $$\rho = \frac{\mathbb{E}[v(t_1)\, v(t_2)]}{\sqrt{\mathbb{E}[(v(t_1))^2]}\sqrt{\mathbb{E}[(v(t_2))^2]}} \qquad \gamma(t_1,t_2) = \gamma(\tau) \qquad \text{(assuming } m=0\text{)}$$

  $$= \frac{\gamma(\tau)}{\gamma(0)}$$

  $\implies |\gamma(\tau)| \leq \gamma(0)$ since $|\rho| \leq 1$ ■

Remark: Normalized covariance function : $\rho(\tau) := \dfrac{\gamma(\tau)}{\gamma(0)}$

$(\implies |\rho(\tau)| \leq 1)$

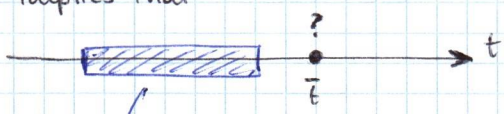
Stochastic process (∞-many random variables) $\longrightarrow$ Stationary process $\longrightarrow$ WHITE NOISE (Unpredictable signal)

UNPREDICTABLE SIGNAL (WHITE NOISE, W.N.)

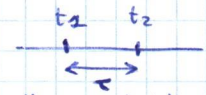
An unpredictable signal is a stationary process with $\gamma(\tau) = 0 \quad \forall \tau \neq 0$.

$$\Longrightarrow \quad \gamma(\tau) = \begin{cases} \lambda^2 & \tau = 0 \\ 0 & \tau \neq 0 \end{cases} \qquad \text{(process variance)}$$

This implies that:



this values are useless to predict the information at the time $\bar{t}$. The past has no use to predict future. Knowing $v(t)$ $t < \bar{t}$ we still have no information to estimate $v(\bar{t})$: WHITNESS PROPERTY



the correlation between the corresponding random variables is null (the past doesn't bring any informations)

Symbols: $\eta \sim WN(0, \lambda^2)$; white noise is with mean $= 0$ and variance λ^2

- $\mathbb{E}[\eta(t)] = 0 \quad \forall t$
- $Var(\eta(t)) = \lambda^2 \quad \forall t$
- $\gamma(\tau) = \begin{cases} \lambda^2 & \tau = 0 \\ 0 & \tau \neq 0 \end{cases}$

We can also consider $\eta(t) \sim WN(m, \lambda^2)$

Example: $v(t) = \alpha \eta(t) + \beta \qquad \alpha, \beta \in \mathbb{R}, \ \eta \sim WN(0, \lambda^2$

find the characteristics of $v(t)$.

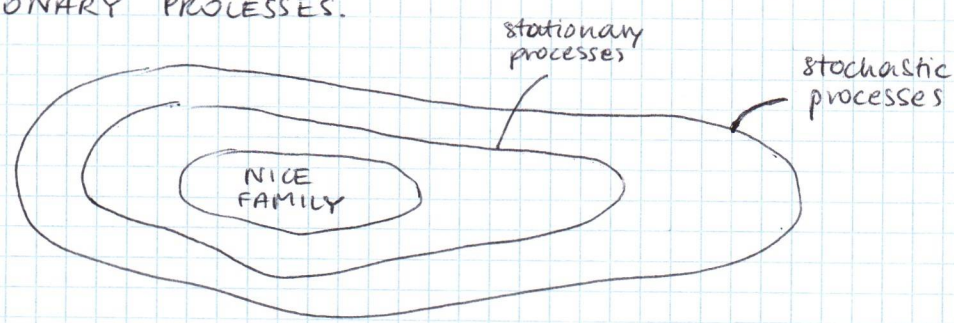
- $\mathbb{E}[v(t)] = \mathbb{E}[\alpha \eta(t) + \beta] = \alpha \mathbb{E}[\eta(t)] + \beta = \beta$
- $Var(v(t)) = \mathbb{E}[(v(t) - \beta)^2] = \mathbb{E}[(\alpha \eta(t))^2] = \alpha^2 \mathbb{E}[(\eta(t))^2]$
  $= \alpha^2 Var(\eta(t)) = \alpha^2 \lambda^2$
- Covariance function:

$$\begin{aligned}
\gamma(t_1, t_2) &= \mathbb{E}[(v(t_1) - \mathbb{E}[v(t_1)])(v(t_2) - \mathbb{E}[v(t_2)])] \\
&= \mathbb{E}[(v(t_1) - \beta)(v(t_2) - \beta)] \\
&= \mathbb{E}[(\alpha \eta(t_1) + \cancel{\beta - \beta})(\alpha \eta(t_2) + \cancel{\beta - \beta})] \\
&= \mathbb{E}[\alpha^2 \eta(t_1) \eta(t_2)] \\
&= \alpha^2 \underbrace{\mathbb{E}[\eta(t_1) \eta(t_2)]}_{\text{covariance function of the WN } (\eta(t))} = \begin{cases} \alpha^2 \lambda^2 & t_1 = t_2 \\ 0 & t_1 \neq t_2 \end{cases} \quad \Big\} \text{ W.N. property}
\end{aligned}$$

$\Longrightarrow$ $v(t)$ is a WN with mean β and variance $\alpha^2 \lambda^2$:

$v(t) \sim WN(\beta, \alpha^2 \lambda^2)$

Starting from the concept of WN, we introduce now a NICE FAMILY of STATIONARY PROCESSES.



MOVING AVERAGE PROCESSES (MA(•))

Consider a white noise WN ($\eta(t)$ s.t. $\eta \sim WN(0, \lambda^2)$) and define a new process $v(t)$ as:

$$v(t) = c_0\, \eta(t) + c_1\, \eta(t-1) \qquad c_0, c_1 \in \mathbb{R} \qquad (MA(1))$$

Find the characteristics of $v(t)$:

- mean value: $\mathbb{E}[v(t)] = c_0\, \mathbb{E}[\eta(t)] + c_1\, \mathbb{E}[\eta(t-1)] = 0$ (constant)
- variance:

$$\begin{aligned}
\mathbb{E}[(v(t)-0)^2] &= \mathbb{E}[(v(t))^2] \\
&= \mathbb{E}[(c_0\, \eta(t) + c_1\, \eta(t-1))^2] \\
&= \mathbb{E}[c_0^2\, \eta^2(t) + c_1^2\, \eta^2(t-1) + 2c_0c_1\, \eta(t)\, \eta(t-1)] \\
&= c_0^2\, \underbrace{\mathbb{E}[\eta^2(t)]}_{\lambda^2} + c_1^2\, \underbrace{\mathbb{E}[\eta^2(t-1)]}_{\lambda^2} + \\
&\quad + 2c_0c_1\, \underbrace{\cancel{\mathbb{E}[\eta(t)\, \eta(t-1)]}}_{\substack{\text{they're uncorrelated so} \\ \mathbb{E}[\eta(t)\eta(t-1)] = \mathbb{E}[\eta(t)]\, \mathbb{E}[\eta(t-1)] \\ \text{and they're both of mean } = 0}} \\
&= (c_0^2 + c_1^2)\, \lambda^2 \quad \text{(constant)}
\end{aligned}$$

- covariance function:

$$\begin{aligned}
\gamma(t, t-1) &= \mathbb{E}[v(t)\, v(t-1)] = \mathbb{E}[(c_0\, \eta(t) + c_1\, \eta(t-1))(c_0\, \eta(t-1) + c_1\, \eta(t-2))] \\
&= c_0^2\, \cancel{\mathbb{E}[\eta(t)\, \eta(t-1)]} + c_0c_1\, \cancel{\mathbb{E}[\eta(t)\, \eta(t-2)]} + c_0c_1\, \underbrace{\mathbb{E}[\eta^2(t-1)]}_{var(\eta(t-1))} + \\
&\quad + c_1^2\, \cancel{\mathbb{E}[\eta(t-1)\, \eta(t-2)]} \\
&= var(\eta(t-1))\, c_0c_1 \\
&= c_0c_1\, \lambda^2 \qquad (\perp\!\!\!\perp t)
\end{aligned}$$

$$\gamma(t, t-2) = [\ldots] = 0$$

$$\gamma(t, t-3) = [\ldots] = 0$$

$$\ldots$$

$\Longrightarrow$ $v(t)$ is a **stationary process**:

- $\mathbb{E}[v(t)] = 0$
- $\gamma(t_1, t_2) = \begin{cases} (c_0^2 + c_1^2)\, \lambda^2 & t_2 = t_1 \\ c_0c_1\, \lambda^2 & t_2 = t_1 \pm 1 \\ 0 & \text{elsewhere} \end{cases}$

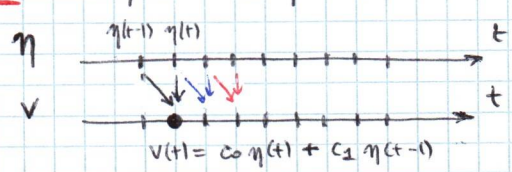
(It's stationary since $\gamma(t_1, t_2)$ depends only on $t_2 - t_1$, i.e. it depends only on how far t_1 and t_2 are)

$\Longrightarrow$ $v(t)$ is a stationary process with null mean value and covariance function $\gamma(\tau)$:

$$\gamma(\tau) = \begin{cases} (c_0^2 + c_1^2)\, \lambda^2 & \tau = 0 \\ c_0c_1\, \lambda^2 & \tau = \pm 1 \\ 0 & \text{elsewhere} \end{cases}$$

Is it a WN? No, because $\gamma(1) \neq 0$ $\Longrightarrow$ $v(t)$ is not a WN

$v(t) = c_0\, \eta(t) + c_1\, \eta(t-1)$ is called MOVING AVERAGE PROCESS of order 1



$MA(1)$: $v(t) = c_0\,\eta(t) + c_1\,\eta(t-1)$

$MA(2)$: $v(t) = c_0\,\eta(t) + c_1\,\eta(t-1) + c_2\,\eta(t-2)$

$MA(n)$: $v(t) = c_0\,\eta(t) + c_1\,\eta(t-1) + c_2\,\eta(t-2) + \dots + c_n\,\eta(t-n)$

$\eta \sim WN(0, \lambda^2)$

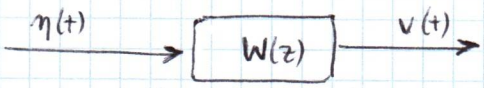
Stationary process s.t. :

- $\mathbb{E}[v(t)] = 0$
- $Var(v(t)) = (c_0^2 + c_1^2 + \dots + c_n^2)\,\lambda^2$
- $$\gamma(\tau) = \begin{cases} (c_0^2 + c_1^2 + \dots + c_n^2)\,\lambda^2 & \tau = 0 \\ (c_0 c_1 + c_1 c_2 + \dots + c_{n-1} c_n)\,\lambda^2 & \tau = 1 \\ (c_0 c_2 + c_1 c_3 + \dots + c_{n-2} c_n)\,\lambda^2 & \tau = 2 \\ \dots & \\ c_0 c_n\,\lambda^2 & \tau = n \\ 0 & \tau > n \end{cases}$$

Remark on the REPRESENTATION OF AN **MA(n)** PROCESS (DYNAMIC REPRESENTATION)

$$\underset{\text{output}}{v(t)} = c_0\,\eta(t) + c_1\,\eta(t-1) + \dots + c_n\,\eta(t-n)$$
$$= c_0\,\eta(t) + c_1 z^{-1}\eta(t) + \dots + c_n z^{-n}\eta(t)$$
$$= \underbrace{(c_0 + c_1 z^{-1} + \dots + c_n z^{-n})}_{:= W(z)}\underset{\text{input}}{\eta(t)}$$

we use the delay op. :
$z^{-1} y(t) = y(t-1)$
$\dots$
$z^{-k} y(t) = y(t-k)$



transfer function from η to v :

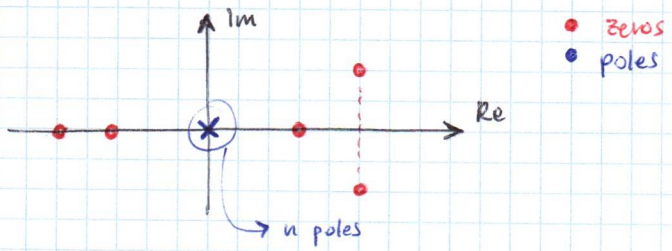
$$W(z) = c_0 + c_1 z^{-1} + \dots + c_n z^{-n} = \frac{c_0 z^n + c_1 z^{n-1} + \dots + c_n}{z^n}$$

considering z as the complex variable we have that $W(z)$ is a rational function (quotient between two polynomials in z)

- poles : complex numbers s.t. the denominator $D(z) = 0$
- zeros : complex numbers s.t. the numerator $N(z) = 0$

general def.

Considering the form of $W(z)$ we have that :



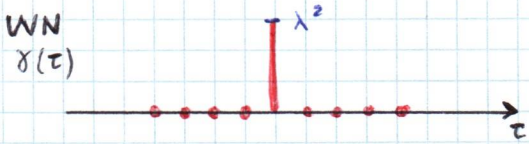
- **poles** of an $MA(n)$ process : $z^n = 0 \Rightarrow$ n poles in the origin of the complex plane
- **zeros** of an $MA(n)$ process : $c_0 z^n + c_1 z^{n-1} + \dots + c_n = 0$ (polynomial with real coefficients $c_0, \dots, c_n$) $\Longrightarrow$ this polynomial has n roots that can be real or complex. If a root is complex then there is another complex root which is the conjugate complex of the previous root.

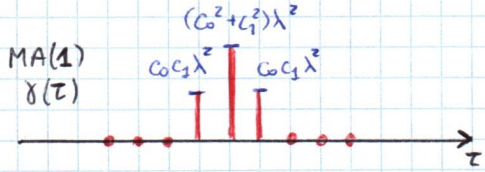
Conclusion : $MA(n)$

An MA process of order n has n poles all located in the origin of the complex plane and n zeros (real or complex) located in various positions of the complex plane.
Poles and zeros are the signature of a signal (so of the process).

WHITE NOISE AND MA(n)

Both white noise and moving average processes are stationary processes, but they're different from each other:





stationary process but not white. (since $\gamma(1) \neq 0$), it's a **colored** noise

(Remark:
- $\gamma(-\tau) = \gamma(\tau)$
- $|\gamma(\tau)| \leq \gamma(0)$
- $\gamma(0) \geq 0$)

MA(n) has longer memory than MA(n-1)

MA(∞)

$$v(t) = c_0\eta(t) + c_1\eta(t-1) + \dots + c_n\eta(t-n) + \dots \qquad \eta \sim WN(0, \lambda^2)$$

Characteristics of $v(t)$:

- mean value: $\mathbb{E}[v(t)] = 0$
- variance of the process: $\gamma(0) = (c_0^2 + c_1^2 + \dots + c_n^2 + \dots)\lambda^2$

  It's an infinite sum, so we have to impose some conditions for its existence:

  Basic condition: $\sum_{i=0}^{+\infty} c_i^2 < +\infty$

  Under this condition, not only the variance of the process $\gamma(0)$ is well defined, the whole covariance function is well defined ($\gamma(\tau)\ \forall\tau$):

  $\gamma(0) = (\sum_i c_i^2)\lambda^2$
  $\gamma(1) = (c_0c_1 + c_1c_2 + \dots)\lambda^2$
  $\gamma(2) = (c_0c_2 + c_1c_3 + \dots)\lambda^2$

Conclusion: under the basic condition, the process MA(∞) is well defined and MA(∞) is a **stationary** process

Recap: WN $\longrightarrow$ MA(1) $\longrightarrow$ MA(n) $\longrightarrow$ MA(∞) (they're all STATIONARY)

AUTOREGRESSIVE PROCESSES (AR(•))

(another family of stationary processes)

AR(1): $v(t) = a\,v(t-1) + \eta(t) \qquad \eta \sim WN(0, \lambda^2)$

autoregressive model of order 1

What type of stochastic process is generated by this model?

We note that: $v(t-1) = a\,v(t-2) + \eta(t-1)$

$v(t-2) = a\,v(t-3) + \eta(t-2)$

$$\Longrightarrow \quad v(t) = a\,v(t-1) + \eta(t)$$
$$= a^2 v(t-2) + a\,\eta(t-1) + \eta(t)$$
$$= a^3 v(t-3) + a^2\eta(t-2) + a\,\eta(t-1) + \eta(t)$$

$$\longrightarrow \quad v(t) = \underbrace{a^k v(t-k)}_{\text{if } |a|<1 \;\Rightarrow\; a^k v(t-k) \to 0 \text{ as } k\to\infty} + \sum_{i=0}^{k} a^i \eta(t-i)$$

In the long term we have:

$$\underline{v(t)} = \eta(t) + a\,\eta(t-1) + a^2\,\eta(t-2) + \dots = \sum_{i=0}^{+\infty} a^i \eta(t-i)$$

solution of the autoregressive model of order 1 (AR(1)) if $|a|<1$

This is an MA(∞) process:

- $c_0 = 1$
- $c_1 = a$
- $c_2 = a^2$
- ...
- $c_k = a^k$

So we can use the MA(∞) formulas:

$$\gamma(0) = (1 + a^2 + a^4 + \dots)\lambda^2 = \frac{1}{1-a^2}\lambda^2 \quad := \text{variance of the process}$$

$$\gamma(1) = (c_0c_1 + c_1c_2 + \dots)\lambda^2 = (1a + aa^2 + \dots)\lambda^2 = a(1 + a^2 + a^4 + \dots)\lambda^2 = a \cdot \frac{1}{1-a^2}\lambda^2$$

$$\gamma(2) = \frac{a^2}{1-a}\lambda^2$$

...

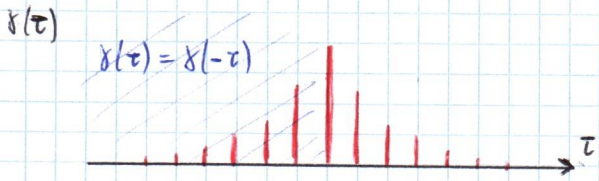
$$\gamma(k) = \frac{a^{|k|}}{1-a}\lambda^2 \qquad (|a|<1\,!)$$

Is this process well defined? Yes, if $(c_0^2 + c_1^2 + \dots) < +\infty$, and we know that:

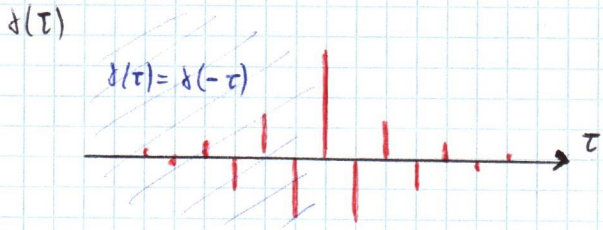
$$\sum_{k=0}^{+\infty} c_k^2 = 1 + a^2 + \dots < +\infty \iff |a| < 1.$$

Conclusion: an AR(1) is a stationary process with the following covariance function:

$$\underline{\gamma(\tau)} = \frac{a^{|\tau|}}{1-a^2}\lambda^2 :$$



$0 < a < 1$



$-1 < a < 0$

Note: an AR process has an infinite memory even if it is of order 1 (AR(1)) (it's very different from an MA(1) process, which has a short memory (only 1 step))

MA(1) vs. AR(1)

$(\eta \sim WN(0, \lambda^2))$

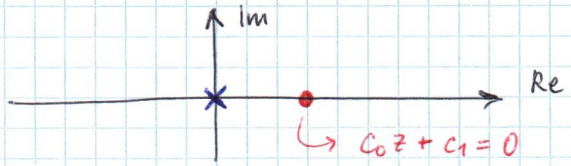
MA(1): $v(t) = c_0\,\eta(t) + c_1\,\eta(t-1)$

$$\gamma(\tau) = \begin{cases} (c_0^2 + c_1^2)\lambda^2 & \tau = 0 \\ (c_0c_1)\lambda^2 & \tau = \pm 1 \\ 0 & |\tau| > 1 \end{cases}$$

Transfer function from η to v:

$$W(z) = c_0 + c_1 z^{-1} \qquad \text{(negative powers notation)}$$

$$= \frac{c_0 z + c_1}{z} \qquad \text{(positive powers notation)}$$



x = pole
• = zero

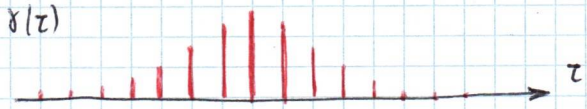
The important characteristic is that the pole is located in the origin.

AR(1):

$$\gamma(\tau) = \begin{cases} \dfrac{1}{1-a^2}\lambda^2 & \tau = 0 \\ \dfrac{a^{|\tau|}}{1-a^2}\lambda^2 & \tau \neq 0 \end{cases} \qquad = a^{|\tau|}\gamma(0)$$

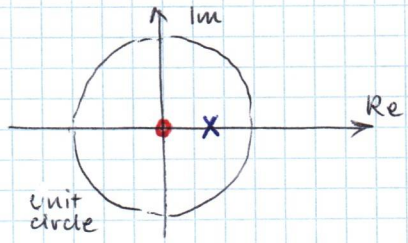
Basic condition: $|a| < 1$

$(0 < a < 1)$



Transfer function from η to v:

$$W(z) = \frac{z}{z-a}$$



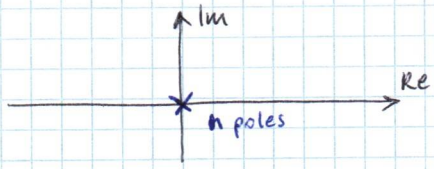
The basic condition $|a| < 1$ is equivalent to saying:
poles must be inside the unit disc
(STABILITY CONDITION)
(the AR(1) generates a stationary process <u>if</u> it is STABLE)

MA(n) vs. AR(n) (generalization)

MA(n): $v(t) = c_0\,\eta(t) + c_1\,\eta(t-1) + \dots + c_n\,\eta(t-n)$

always STABLE, it always generates a stationary process

$$W(z) = \frac{c_0 z^n + \dots + c_n}{z^n}$$



All the poles are in the origin, so they're all inside the unit disc
$\Longrightarrow$ the system is always **STABLE**

AR(n): $v(t) = a_1 v(t-1) + a_2 v(t-2) + \dots + a_n v(t-n) + \eta(t)$

generates a stationary process <u>only</u> under the condition of stability

$$W(z) = \frac{z^n}{z^n - a_1 z^{n-1} - \dots - a_n}$$

The poles position in the complex plane depends on the value of the coefficients $a_1, \dots, a_n$. The condition for the STABILITY is that all the poles must be inside the unit circle in the complex plane. If the system is stable, then the autoregressive model generates a stationary process.

ARMA PROCESSES (combination of MA and AR)

$\eta \sim WN(0, \lambda^2)$

$$v(t) := \underbrace{c_0 \eta(t) + c_1 \eta(t-1) + \dots + c_n \eta(t-n)}_{MA(n_c)} + \underbrace{a_1 v(t-1) + a_2 v(t-2) + \dots + a_n v(t-n)}_{AR(n_a)}$$

$\Longrightarrow$ ARMA(n_a, n_c) :

$$\underset{WN}{\eta(t)} \longrightarrow \boxed{W(z)} \xrightarrow[\text{process}]{v(t)}$$

The process is stationary under the condition of stability (i.e. all poles of $W(z)$ are located in the unit disc of the complex plane).

Before we procede we want to discuss more on the **COMPUTATION OF $\gamma(\tau)$ IN AR(1)**

We saw that an AR(1) process is equivalent to an MA(∞) process, and then we used the formulas for $\gamma(\tau)$ of the MA(∞) process.
There is an alternative way of computing $\gamma(\tau)$ for AR(1) : the direct way. This will lead to the YULE-WALKER EQUATION.

AR(1) : $v(t) = a\, v(t-1) + \eta(t)$

Compute first the variance $\gamma(0)$:

$$\gamma(0) = Var(v(t)) = \mathbb{E}[(v(t) - 0)^2] = \mathbb{E}[v^2(t)]$$

$$= \mathbb{E}[(a\, v(t-1) + \eta(t))^2]$$

$$= a^2 \underbrace{\mathbb{E}[v^2(t-1)]}_{\substack{\text{variance of the process at time } t-1, \\ \text{but the process is stationary so the} \\ \text{variance is the same} \\ \Longrightarrow \gamma(0)}} + \underbrace{\mathbb{E}[\eta^2(t)]}_{\substack{Var(\eta(t)) \\ = \lambda^2}} + 2a\, \underbrace{\cancel{\mathbb{E}[v(t-1)\, \eta(t)]}}_{\substack{v(t-1) \text{ is a function of the past of } \eta(\cdot) \\ \text{up to } t-1 \quad (v(t-1) = a\, v(t-2) + \eta(t-1)) \\ \Longrightarrow v(t-1) = f(\eta(t-1), \eta(t-2), \dots) \\ \Longrightarrow v(t-1) \perp\!\!\!\perp \eta(t) \\ \Longrightarrow \mathbb{E}[v(t-1)\, \eta(t)] = \underbrace{\cancel{\mathbb{E}[v(t-1)]}}_{=0}\, \underbrace{\cancel{\mathbb{E}[\eta(t)]}}_{=0} = 0}}$$

$$= a^2 \gamma(0) + \lambda^2$$

$$\Longrightarrow \gamma(0) = a^2 \gamma(0) + \lambda^2 \Longrightarrow \gamma(0) = \frac{1}{1-a^2} \lambda^2$$

We want to compute $\gamma(\tau)$:

$$v(t) = a\, v(t-1) + \eta(t) \xrightarrow{\cdot\, v(t-\tau)} v(t)\, v(t-\tau) = a\, v(t-1)\, v(t-\tau) + \eta(t)\, v(t-\tau)$$

$$\xrightarrow{\mathbb{E}[\cdot]} \underbrace{\mathbb{E}[v(t)\, v(t-\tau)]}_{\gamma(\tau)} = \underbrace{\mathbb{E}[a\, v(t-1)\, v(t-\tau)]}_{a\, \gamma(\tau-1)} + \underbrace{\cancel{\mathbb{E}[\eta(t)\, v(t-\tau)]}}_{\substack{v(t-\tau) = f(\eta(t-\tau), \eta(t-\tau-1), \dots) \\ \text{(function of } \eta(\cdot) \text{ up to } t-\tau), \ \eta(t) \\ \text{is a WN, so is a function only of } t \\ \Longrightarrow \eta(t) \perp\!\!\!\perp v(t-\tau)}}$$

$$\gamma(\tau) = \begin{cases} \dfrac{1}{1-a^2} \lambda^2 & \tau = 0 \\ a\, \gamma(\tau-1) & \tau > 0 \end{cases}$$

YULE-WALKER EQUATIONS
(equations if we consider $\gamma(0)$ and $\gamma(\tau), |\tau| > 0$)

FREQUENCY DOMAIN

We have seen that in a time domain a stochastic process is represented with:

- mean value m
- variance $\gamma(0)$
- covariance function $\gamma(\tau)$ (we assume "stationary" processes)

SPECTRAL REPRESENTATION

We define **SPECTRUM** of a stationary process the Fourier's transform of the covariance function $\gamma(\tau)$:

$$\boxed{\underbrace{\Gamma(\omega)}_{\text{spectrum}} := \mathcal{F}[\underbrace{\gamma(\tau)}_{\text{covariance}}] = \sum_{\tau=-\infty}^{+\infty} \gamma(\tau)\, e^{-j\omega\tau}}$$

ω is a real variable.

$$\Longrightarrow \Gamma(\omega) = \dots + \gamma(-2)e^{+j\omega 2} + \gamma(-1)e^{+j\omega} + \gamma(0) + \gamma(1)e^{-j\omega} + \gamma(2)e^{-j\omega 2} + \dots$$

Since $\gamma(\tau) = \gamma(-\tau)$ we can couple the terms:

$$\gamma(k)e^{-j\omega k} + \gamma(-k)e^{+j\omega k} = \gamma(k)\left(e^{-j\omega k} + e^{+j\omega k}\right)$$

$$\Longrightarrow \Gamma(\omega) = \gamma(0) + \gamma(1)\left(e^{-j\omega} + e^{+j\omega}\right) + \gamma(2)\left(e^{-j\omega 2} + e^{+j\omega 2}\right) + \dots$$

EULER FORMULA: $\cos(k\omega) = \frac{1}{2}\left(e^{-j(k\omega)} + e^{+j(k\omega)}\right)$

$$= \gamma(0) + \gamma(1)\left(2\cos(\omega)\right) + \gamma(2)\left(2\cos(2\omega)\right) + \gamma(3)\left(2\cos(3\omega)\right) + \dots$$

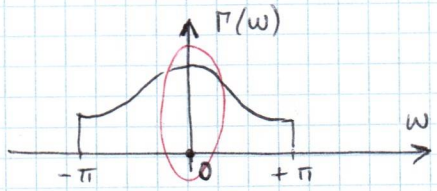
$\Longrightarrow$ $\Gamma(\omega)$ is <u>real</u> (since it's the Σ of real functions) even if in the first expression it seems to be complex. ($\Gamma(\omega)$ real function of a real variable ω)

$\Longrightarrow$ $\Gamma(\omega)$ is <u>periodic</u> with period 2π

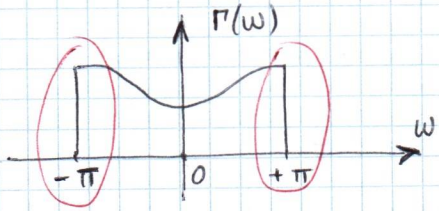
$\Longrightarrow$ $\Gamma(\omega)$ is <u>even</u>: $\Gamma(-\omega) = \Gamma(+\omega)$

How do we represent the spectrum? Representations:

1. $\Gamma(\omega)$ from $\omega = -\pi$ to $\omega = +\pi$ (since it's periodic)

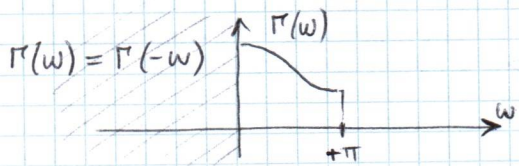


low frequencies are dominating



high frequencies are dominating

2. $\Gamma(\omega)$ from $\omega = 0$ to $\omega = +\pi$ (since it's even):

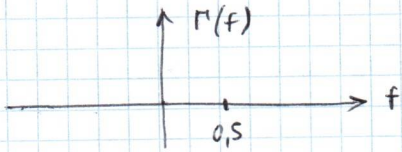


3. ω angular frequency

$\Longrightarrow$ $\omega = 2\pi f$, f: frequency

$\omega = \pi \Longrightarrow f = 0.5$

$\Longrightarrow$



Another property of the spectrum: $\underline{\Gamma(\omega) \geq 0.}$

Example: **WHITE NOISE SPECTRUM**:

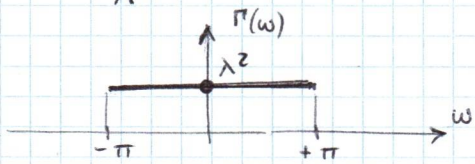
$v(t) \sim WN(0, \lambda^2)$

$$\gamma(\tau) = \begin{cases} \lambda^2 & \tau = 0 \\ 0 & \tau \neq 0 \end{cases}$$

Spectrum:

$$\Gamma(\omega) = \gamma(0) + \cancel{\gamma(1)\left(e^{-j\omega} + e^{+j\omega}\right) + \gamma(2)\left(e^{-j\omega 2} + e^{+j\omega 2}\right) + \dots}$$

$$= \lambda^2$$



$\Longrightarrow$ the spectrum is constant

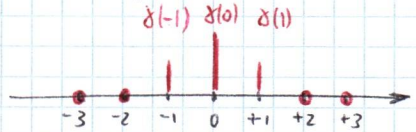
$\Longrightarrow$ all frequencies provide the same contribution to the variability of the signal (there is no dominance of low/high frequencies, all frequencies are important)

(MAIN WN property!)

Example: **MA(1) PROCESS SPECTRUM**:

$v(t) = c_0\, \eta(t) + c_1\, \eta(t-1)$ $\equiv$ linear combination of two samples of WN ($\eta \sim WN(0, \lambda^2)$)

$$\gamma(\tau) = \begin{cases} (c_0^2 + c_1^2)\lambda^2 & \tau = 0 \\ c_0 c_1 \lambda^2 & |\tau| = 1 \\ 0 & |\tau| > 1 \end{cases}$$



$(c_0 c_1 > 0)$ ⊛

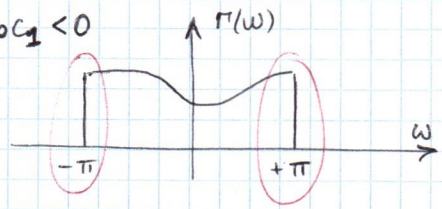
Spectrum:

$$\Gamma(\omega) = \gamma(0) + \gamma(1)\left(e^{-j\omega} + e^{+j\omega}\right) + \cancel{\gamma(2)\left(e^{-j\omega 2} + e^{+j\omega 2}\right) + \dots}$$

$$= (c_0^2 + c_1^2)\lambda^2 + c_0 c_1 \lambda^2 \left(e^{j\omega} + e^{-j\omega}\right)$$

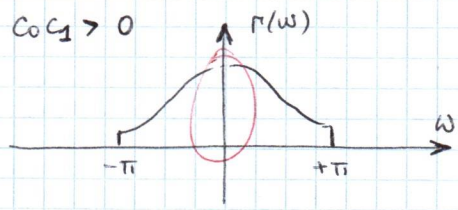
$$= \lambda^2 \left(c_0^2 + c_1^2 + 2 c_0 c_1 \cos(\omega)\right)$$

$c_0 c_1 < 0$



high frequencies are dominating (the realization will be quicky changing (up and down))

$c_0 c_1 > 0$



low frequencies are dominating (the signal moves very slowly)

⊛ "little/small colored noise" since $\gamma \neq 0$ only for $|\tau| = 1$

ANTI - TRANSFORM FORMULA

$$\Gamma(\omega) = \sum_{\tau=-\infty}^{+\infty} \gamma(\tau) e^{-j\omega\tau}$$

$\gamma(\tau) \xrightarrow{\text{FOURIER'S TR.}} \Gamma(\omega)$

function in the time domain → function in the frequence domain

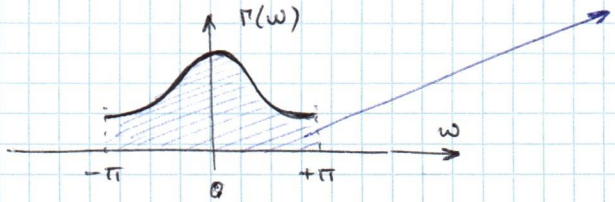
(τ discrete variable → ω continuous variable)

How to come back? Using the anti-transform function.

ANTI- FOURIER TRANSFORM :

$$\gamma(\tau) := \frac{1}{2\pi} \int_{-\pi}^{+\pi} \Gamma(\omega)\, e^{+j\omega\tau} \, d\omega$$

Note that, letting $\tau = 0$: $\gamma(0) = \frac{1}{2\pi} \boxed{\int_{-\pi}^{\pi} \Gamma(\omega)\, d\omega}$



⟹ up to a coefficient $\frac{1}{2\pi}$, the area below the spectrum from $\omega = -\pi$ to $\omega = +\pi$ is the variance of the process.

COMPLEX SPECTRUM

We introduce the complex spectrum : $\boxed{\phi(z) = \sum_{\tau=-\infty}^{+\infty} \gamma(\tau)\, z^{-\tau}}$

function of a complex variable z

We can obtain the real spectrum $\Gamma(\omega)$ from $\phi(z)$:

$$\Gamma(\omega) = \sum_{\tau=-\infty}^{+\infty} \gamma(\tau)\, e^{-j\omega\tau} = \phi(e^{j\omega})$$

COMPUTATION OF THE SPECTRUM

Consider the <u>ARMA process</u> : $\xrightarrow{\eta \sim WN} \boxed{W(z)} \xrightarrow{v}$.

We want to find the spectrum of $v(\cdot)$. How can we procede?
(We assume the system to be stable ⟹ v is stationary)

1. find the covariance function of v
2. perform the Fourier's transform

This is easy if $W(z)$ is simple. If it's more complex then the computation of the covariance function could be difficult ⟹ we need an other way.

Consider a <u>RATIONAL process</u> : $\xrightarrow[\text{stationary process}]{u} \boxed{W(z)} \xrightarrow{y}$.
(more complex case)

We assume $W(z)$ stable ⟹ y is stationary. We want to find the spectrum of y knowing the characteristics of the input process u.
We should start from the covariance function of u, calculate the covariance function of y and then apply the Fourier's transform. It's a long way.

⟹ ∃ an other way for these two cases:

FUNDAMENTAL THEOREM OF SPECTRUM ANALYSIS

(It relates the spectrum of the output with the spectrum of the input)

Consider the ARMA system :

$$\phi_{vv}(z) = \sum_{\tau=-\infty}^{+\infty} \gamma(\tau)\, e^{-j\omega\tau}$$

complex spectrum of the output

$\xrightarrow{\eta \sim WN(0,\lambda^2)} \boxed{W(z)} \xrightarrow{v}$

(We know that the spectrum of $\eta \sim WN(0,\lambda^2)$ is constant and $= \lambda^2$)

$$\Longrightarrow \quad \boxed{\Phi_{vv}(z) = W(z)\, W(z^{-1})\, \lambda^2}$$

:= MAGIC FORMULA I
in complex variable z
(enable to compute the spectrum of the output of an ARMA system without passing through the covariance function)

$$\Longrightarrow \quad \boxed{\Gamma_{vv}(\omega) = |W(e^{j\omega})|^2 \lambda^2}$$

:= MAGIC FORMULA II
in real variable ω

indeed :

$$\Gamma_{vv}(\omega) = \Phi_{vv}(z)\big|_{z=e^{j\omega}}$$

$$\Longrightarrow \Phi_{vv}(z)\big|_{z=e^{j\omega}} = W(e^{j\omega})\, W(e^{-j\omega})\, \lambda^2$$
$$= W(e^{j\omega})\, \overline{W(e^{+j\omega})}\, \lambda^2$$
$$= |W(e^{j\omega})|^2 \lambda^2 \qquad \blacksquare$$

Example : MA process: $v(t) = \eta(t) + \eta(t-1)$ $\qquad (c_0 = c_1 = 1)$

Compute the spectrum of v :

(1) from the definition of spectrum
(2) from the fundamental theorem of spectral analysis (MAGIC FORMULA)

---

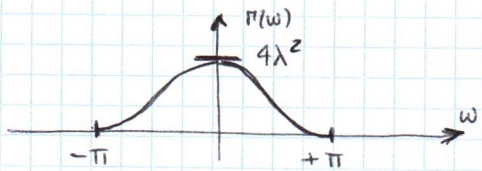
(1) Covariance function of $v(\cdot)$:

$$\gamma(\tau) = \begin{cases} (c_0^2 + c_1^2)\lambda^2 & \tau = 0 \\ c_0 c_1 \lambda^2 & |\tau| = 1 \\ 0 & |\tau| > 1 \end{cases}$$

Spectrum :

$$\Phi_{vv}(z) = \sum_{\tau=-\infty}^{+\infty} \gamma(\tau)\, z^{-\tau}$$
$$= \gamma(0) + \gamma(1)\, z^{-1} + \gamma(-1)\, z^{+1}$$
$$= (c_0^2 + c_1^2)\lambda^2 + c_0 c_1 \lambda^2 z^{-1} + c_0 c_1 \lambda^2 z^{+1}$$
$$= 2\lambda^2 + \lambda^2 z^{-1} + \lambda^2 z^{+1}$$
$$= (2 + z^{-1} + z^{+1})\lambda^2$$

$$\Gamma_{vv}(\omega) = \Phi(z)\big|_{z=e^{j\omega}} = (2 + 2\cos(\omega))\lambda^2$$



---

(2) Fund. thm. : $\Phi_{vv}(z) = W(z)\, W(z^{-1})\, \lambda^2$

$$v(t) = \eta(t) + \eta(t-1) = \eta(t) + z^{-1}\eta(t) = (1 + z^{-1})\,\eta(t)$$

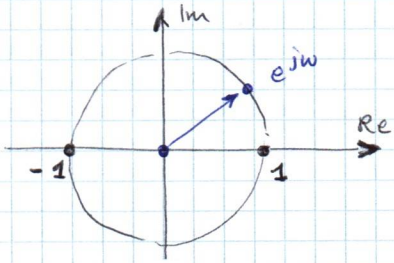
$$W(z) = 1 - z^{-1} = \frac{z+1}{z}$$

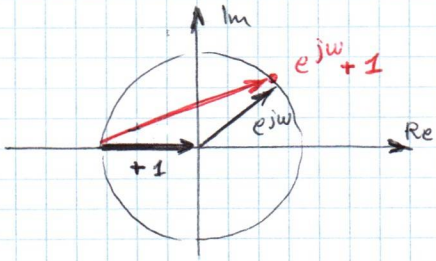
$$\Longrightarrow \Phi_{vv}(z) = (1 + z^{-1})(1 + z)\lambda^2$$
$$= (2 + z + z^{-1})\lambda^2$$

Considerations on the drawing :

$$\Gamma(\omega) = \left| \frac{e^{j\omega} + 1}{e^{j\omega}} \right|^2 \lambda^2 \qquad \begin{array}{l} \to V_n(\omega) \\ \to V_d(\omega) \end{array}$$

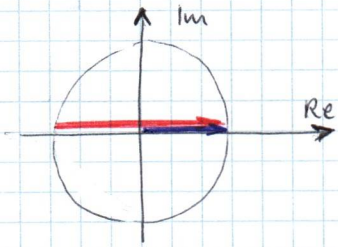
$V_n(\omega)$ and $V_d(\omega)$ are complex vector functions of ω :





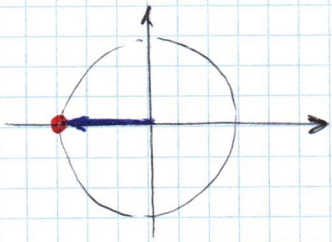
What happens when ω goes from $-\pi$ to $+\pi$?

$\omega = 0$:



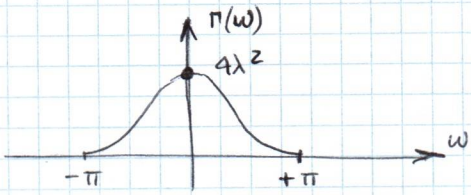
$$\left.\begin{array}{l} V_n(0) = 2 \\ V_d(0) = 1 \end{array}\right\} \Longrightarrow \Gamma(0) = 4\lambda^2$$

$\omega = \pi$:



$$\left.\begin{array}{l} V_n(\pi) = 0 \\ V_d(\pi) = -1 \end{array}\right\} \Longrightarrow \Gamma(\pi) = 0$$

$\Longrightarrow$



Low frequencies are dominating, the realizations will change very slowly

Example : AR(2) : $v(t) = -0{,}81\, v(t-2) + \eta(t)$ $\qquad \eta(t) \sim WN(0, \lambda^2)$

Compute the transfer function from η to v and by appling the fundamental theorem of spectr. analysis find $\Gamma(\omega)$ and draw it from $\omega = -\pi$ to $\omega = +\pi$

$$v(t) = \frac{1}{1 + 0{,}81\, z^{-2}}\, \eta(t) = \boxed{\frac{z^2}{z^2 + 0{,}81}}\, \eta(t)$$

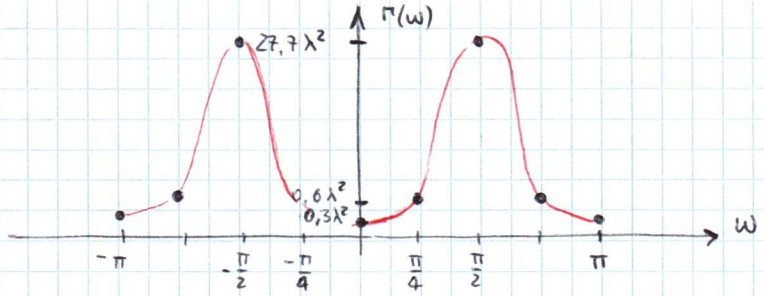
$W(z)$ $(z^2 + 0{,}81 = 0$ $\Longrightarrow z = \pm i\, 0{,}9 \Longrightarrow$ the system is stable

$$|W(e^{j\omega})| = \left| \frac{e^{2j\omega}}{e^{2j\omega} + 0{,}81} \right| = \frac{|e^{2j\omega}|}{|e^{2j\omega} + 0{,}81|} = \frac{1}{|e^{2j\omega} + 0{,}81|}$$

$$|W(e^{j\omega})|^2 = \frac{1}{|\cos(2\omega) + i\sin(2\omega) + 0{,}81|^2} = \frac{1}{(\cos(2\omega) + 0{,}81)^2 + (\sin(2\omega))^2}$$

$$\Gamma_{vv}(\omega) = \frac{\lambda^2}{(\cos(2\omega) + 0{,}81)^2 + (\sin(2\omega))^2}$$

- $\Gamma_{vv}(0) = \Gamma_{vv}(\pi) = \dfrac{\lambda^2}{(1{,}81)^2} = 0{,}3\,\lambda^2$
- $\Gamma_{vv}\left(\frac{\pi}{4}\right) = \Gamma_{vv}\left(\frac{3}{4}\pi\right) = 0{,}6\,\lambda^2$
- $\Gamma_{vv}\left(\frac{\pi}{2}\right) = 27{,}7\,\lambda^2$



Generalization: we stated the fundamental theorem of spectrum analysis in the case when the input is a WN:

$$y\,(WN) \longrightarrow \boxed{W(z)} \xrightarrow{\ v\ }$$

but the theorem holds even if the input is any stationary process u:

$$u \longrightarrow \boxed{W(z)} \xrightarrow{\ y\ }$$

: $W(z)$ stable system
u stationary process with spectrum $\Phi_{uu}(z)$

$$\Longrightarrow \quad \boxed{\Phi_{yy}(z) = W(z)\,W(z^{-1})\,\underbrace{\Phi_{uu}(z)}_{\text{spectrum of the input}}} \quad \longleftarrow$$

valid for any input u (stationary process) and any $W(z)$

Moreover, it holds even in a multivariable (≡multidimensional) case, when u and y are vectors. (In this case, instead of a single transfer function $W(z)$ we use a matrix transfer function from the various inputs to the various outputs).

Observation:

$$u \longrightarrow \boxed{W(z)} \xrightarrow{\ y\ }$$

Frequency response: if u is a sinusoid of angular frequency $\bar{\omega}$ and amplitude U:

$$u(t) = U \sin(\bar{\omega} t)$$

under the assumptions that the system is stable $\Longrightarrow$

$$\begin{cases} y(t) = Y \sin(\bar{\omega} t + \alpha) \\ Y = U\,\underbrace{|W(z)|\big|_{z=e^{j\bar{\omega}}}} \end{cases}$$

this transfer function evaluated in $e^{j\bar{\omega}}$ is the **GAIN** of the system

---

(Back to the theorem): $u \longrightarrow \boxed{W(z)} \xrightarrow{\ y\ }$

- $u \sim WN(0, \lambda^2) \Longrightarrow \Gamma(\omega) = \underbrace{|W(z)|\big|^2_{z=e^{j\omega}}}\,\lambda^2 \quad \geq 0$

  (by the way, we can see this as the gain)

- $u \sim$ any (stochastic) stationary process with spectrum $\Gamma_{uu}(\omega)$
  $\Longrightarrow$ y is a stationary process with the spectrum:

$$\boxed{\underbrace{\Gamma_{yy}(\omega)}_{\geq 0} = |W(z)|^2\big|_{z=e^{j\omega}}\ \Gamma_{uu}(\omega)}$$

:= Fundamental theorem of spectral analysis

Remark. Complex spectra:

- $\Phi_{yy}(z) = W(z)\,W(z^{-1})\,\lambda^2 \qquad u \sim WN$
- $\Phi_{yy}(z) = W(z)\,W(z^{-1})\,\Phi_{uu}(z) \qquad u \sim \forall$

OBSERVATION ON ARMA PROCESSES : ARMA (n_a, n_c)

$$v(t) = \underbrace{\boxed{a_1 v(t-1) + a_2 v(t-2) + \dots + a_{n_a} v(t-n_a)}}_{AR(n_a)} + \underbrace{\boxed{c_0 \eta(t) + c_1 \eta(t-1) + \dots + c_{n_c} \eta(t-n_c)}}_{MA(n_c)}$$

n_a = degree of the AR part

n_c = degree of the MA part

In operatorial norm:

$$C(z) = c_0 + c_1 z^{-1} + \dots + c_{n_c} z^{-n_c}$$

$$A(z) = 1 - a_1 z^{-1} - \dots - a_{n_a} z^{-n_a}$$

$\longrightarrow$ We can write the ARMA model as: $A(z)\, v(t) = C(z)\, \eta(t)$, $\eta \sim WN(0, \lambda^2)$

Corresponding transfer function:

$$W(z) = \frac{C(z)}{A(z)} = \frac{c_0 + c_1 z^{-1} + \dots + c_{n_c} z^{-n_c}}{1 - a_1 z^{-1} - \dots - a_{n_a} z^{-n_a}} \cdot \frac{z^n}{z^n}, \quad n = \max(n_a, n_c)$$

$$= \frac{c_0 z^n + c_1 z^{n-1} + \dots + c_{n_c} z^{n-n_c}}{z^n - a_1 z^{n-1} - \dots - a_{n_a} z^{n-n_a}}$$

Observation: any ARMA process generates a MA(∞) process well defined (under the stability assumptions). Stated in an other way: if we assume that $W(z)$ is the transfer function of a stable system (i.e. all poles are inside the unit disc), then the ARMA process is equivalent to an MA(∞) process well defined.

So, $W(z)$ can always be written as:

$$W(z) = w_0 + w_1 z^{-1} + w_2 z^{-2} + \dots \longrightarrow +\infty$$

MA(∞) expression of an ARMA process

How can we derive this expression from the transfer function? By the **LONG DIVISION** algorithm.

Example: AR(1): $v(t) = a\, v(t-1) + \eta(t)$

$$W(z) = \frac{1}{1 - a z^{-1}} = \frac{N(z)}{D(z)}$$

Derivation of the MA(∞) expression with long division: $N(z) \,\big|\, D(z)$:

$$\begin{array}{l|l}
1 & 1 - a z^{-1} \\
\hline
1 - a z^{-1} & 1 + a z^{-1} + a^2 z^{-2} + \dots \longrightarrow \infty \\
\hline
\cancel{1} \quad a z^{-1} & \\
a z^{-1} - a^2 z^{-2} & \\
\hline
\cancel{a z^{-1}} \quad a^2 z^{-2} & \\
a^2 z^{-2} - a^3 z^{-3} & \\
\hline
\cancel{a^2 z^{-2}} \quad \dots &
\end{array}$$

$\infty \downarrow$

$$\Longrightarrow \frac{1}{1 - a z^{-1}} = \underbrace{1}_{w_0} + \underbrace{a}_{w_1} z^{-1} + \underbrace{a^2}_{w_2} z^{-2} + \underbrace{a^3}_{w_3} z^{-3} + \dots$$

= coeff. of MA(∞) represent. of the ARMA process

ARMA representation

- $v(t) = \dfrac{1}{1 - a z^{-1}} \eta(t)$

MA(∞) representation

- $v(t) = w_0 \eta(t) + w_1 \eta(t-1) + w_2 \eta(t-2) + \dots \longrightarrow \infty$

where: $\begin{cases} w_0 = 1 \\ w_1 = a \\ w_2 = a^2 \\ \downarrow \infty \end{cases}$

Note that this can be done for any ARMA process, not only this example.
Moreover, note that in the first representation (ARMA'S) we mix time with the operator z.

Interpretation of the coefficients w_i

$$v(t) = w_0\,\eta(t) + w_1\,\eta(t-1) + w_2\,\eta(t-2) + \dots$$

For the moment let's assume that η is not a WN, suppose it's an IMPULSE:

$$\eta(t) = \begin{cases} 1 & t = 0 \\ 0 & t \neq 0 \end{cases} \implies v(t)\ ?$$

(what is the output?)

$$\implies v(0) = w_0\,\eta(0) + \cancel{w_1\,\eta(-1)} + \cancel{w_2\,\eta(-2)} + \dots = w_0$$

$$v(1) = \cancel{w_0\,\eta(1)} + w_1\,\eta(0) + \cancel{w_2\,\eta(-1)} + \dots = w_1$$

$$[\dots]$$

$$v(k) = w_k$$

Conclusion: in the MA(∞) expression of an ARMA process, i.e. $v(t) = w_0\eta(t) + w_1\eta(t-1) + \dots$, the coefficients $w_0, w_1, w_2, \dots$ are the impulse response of $W(z)$.

Reminder: for a stable system (i.e. all poles of $W(z)$ inside the unit disc) we have that: $w_i \longrightarrow 0$ as $i \to \infty$, more precisely it goes to 0 in the following sense: $w_0^2 + w_1^2 + w_2^2 + \dots < +\infty$

$\implies$ the MA(∞) representation defines a well posed stationary process.

MULTIPLICITY OF ARMA REPRESENTATION

(for a given stationary problem)

Suppose that we have a stationary process $v(t)$. We can characterize it in many ways:

- mean value + covariance function
- mean value + spectrum
- dynamical representation ($\eta \to \boxed{W(z)} \xrightarrow{v}$)
  (we can represent the process as the output of a dynamical system fed by white noise)

What is the link between these representations?
MAIN LINK: relationship between spectrum and dynamical representation:
(main theorem of spectral analysis)

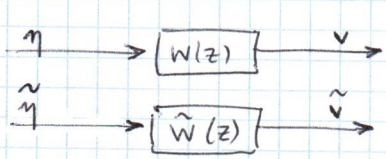
$$\phi(z) = W(z)\,W(z^{-1})\,\lambda^2$$

variance of the WN

ARMA REPRESENTATION
(≡ DYNAMICAL REPR.)

Is the ARMA representation ($W(z)$) unique? Are there more representations with the same spectrum? There are infinitely many ARMA representations for the same stationary process. (There is a multiplicity of the ARMA repr. of the same process).
There are 4 reasons for multiplicity.

1. RESCALING EFFECT
   Consider:



Suppose they're related like:

$$\tilde{W}(z) = \frac{1}{\alpha}\,W(z), \quad \tilde{\lambda}^2 = \alpha^2\lambda^2$$

$$(\eta \sim WN(0,\lambda^2),\ \tilde{\eta} \sim WN(0,\tilde{\lambda}^2))$$

1. We claim that $v = \tilde{v}$ (same spectrum, same covariance function)

$$\tilde{\Phi}(z) = \tilde{W}(z)\,\tilde{W}(z^{-1})\,\tilde{\lambda}^2$$

complex spectrum of $\tilde{v}$

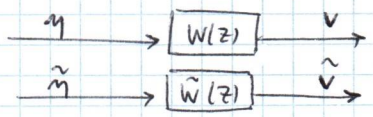
$$= \frac{1}{\alpha} W(z)\,\frac{1}{\alpha} W(z^{-1})\,\alpha^2\,\lambda^2$$

$$= W(z)\,W(z^{-1})\,\lambda^2$$

$$= \Phi(z)$$ complex spectrum of v

2. DELAY EFFECT

Consider:



Suppose: $\underbrace{\tilde{W}(z) = z^{-k} W(z)}$, $\tilde{\lambda}^2 = \lambda^2$

z^{-k} is a k-steps delay;
$\tilde{v}(t) = v(t-k)$: the output of the new system is the previous output delayed for k steps.
If we have a stationary process (stochastic quantity defined from $-\infty$ to $+\infty$) and we translate it of k steps what changes? Nothing, because it's stationary and its spectral / covariance characteristics don't change up to translations. The delay effect doesn't change the essence of the process.

We claim that $v = \tilde{v}$.

$$\tilde{\Phi}(z) = \tilde{W}(z)\,\tilde{W}(z^{-1})\,\tilde{\lambda}^2$$

$$= \cancel{z^{-k}}\, W(z)\, \cancel{z^{+k}}\, W(z)\,\lambda^2$$

$$= \Phi(z)$$

3. COMMON FACTOR EFFECT

$$\tilde{W}(z) = W(z)\,\frac{z+\alpha}{z+\alpha}$$

4. ALL-PASS FILTER

Consider:

$$T(z) = \rho\,\underbrace{\frac{z+\alpha}{z+\frac{1}{\alpha}}}_{\text{the zero is the reciprocal of the pole}} \quad : \quad \xrightarrow[WN]{\eta} \boxed{T(z)} \xrightarrow{v} \quad \Longrightarrow \quad v\,?$$

$$\Phi(z) = T(z)\,T(z^{-1})\,\lambda^2 = \rho\,\frac{z+\alpha}{z+\frac{1}{\alpha}}\,\frac{\frac{1}{z}+\alpha}{\frac{1}{z}+\frac{1}{\alpha}}\,\lambda^2 = \rho\,\lambda^2\,\alpha^2$$

$\Longrightarrow$ taking $\rho = \frac{1}{\alpha}$ we have $\Phi(z) = \lambda^2$

$\Longrightarrow$ the output is a WN as well (with the same variance of the input), this $T(z)$ is very special: if we have a WN as an input then we have a WN as an output ($\xrightarrow{WN} \boxed{T(z)} \xrightarrow{WN}$), this transfer function doesn't modify the spectrum. This $T(z)$ is called ALL-PASS FILTER.

In general, for any input we have:

$$\tilde{W}(z) = W(z)\,T(z) \Longrightarrow \tilde{\Phi}(z) = W(z)\,\underbrace{\boxed{T(z)\,T(z^{-1})}}_{=1 \text{ if } \rho = \frac{1}{\alpha}}\,W(z^{-1})\,\lambda^2$$

$$= \Phi(z) \Longleftarrow \text{unchanged spectrum}$$

$\Longrightarrow$ For a given stationary process we have seen that there are infinitely many ARMA representations. What do we do? We take any ARMA representation? If our goal is to make a prediction (solve the prediction problem) then it is advisable to choose a very special ARMA representation among the ∞-tely many. There is a "nice" representation.

NICE REPRESENTATION OF A STATIONARY PROCESS: **CANONICAL REPR.** $(\hat{W}(z))$

(for preventing the molteplicity) (we'll characterize it with $\wedge$)

1. $W(z)$ has monic numerator and denominator (i.e. the coefficient of the heighest power is 1). (ex. $v(t) = c_0\eta(t) + c_1\eta(t-1) \Rightarrow \tilde{v}(t) = \eta(t) + c_1\eta(t-1)$)
2. Numerator and denominator have the same degree
3. Numerator and denominator have no common factor
4. All poles and zeros of $W(z)$ are inside the unit disk

$\downarrow$

they're avoiding the multiplicity factors defined with the same number (i.e. "monic" avoids the rescaling effect ...)

With this ways to avoid the multiplicity we define the **canonical representation**:

SPECTRAL FACTORIZATION THEOREM

Consider a stationary ARMA process. Then there exist an unique representation satisfying the following conditions:

1. Numerator and denominator:
   - are monic (no rescaling)
   - have the same degree (no delay)
   - are coprime (there is no common factor) (no common factor)
2. poles and zeros of $W(z)$ are inside the unit disk (no hidden ALL-PASS filter)

Example: $\eta \longrightarrow \boxed{W(z)} \xrightarrow{v}$ $\qquad \eta \sim WN(0, \lambda^2)$

$v(t) = 2\eta(t) + \eta(t-1)$ $\qquad$ ($MA(z)$: $c_0 = 2$, $c_1 = 1$)

Is this a canonical representation? **No.**

$W(z) = 2 + z^{-1} \Longrightarrow$ not monic $\left(2 + \frac{1}{z} = \frac{2z+1}{z}\right)$

$\Longrightarrow v(t) = \tilde{\eta}(t) + \frac{1}{2}\tilde{\eta}(t-1)$ $\qquad$ ($\tilde{\eta}(t) = 2\eta(t)$, $\tilde{\eta} \sim WN(0, 4\lambda^2)$)

canonical representation

Exercise (for us): $W(z) = 2z^{-1} + 4z^{-2}$ is not canonical, find the correspondent canonical representation.

Observation: we're working with discrete time (t is an integer). When we draw $\Gamma(\omega)$ we draw it from $-\pi$ to $+\pi$ so the maximum angular frequency is $\omega_{max} = \pi$. The corresponding maximum frequency is $f_{max} = 0{,}5$. It's not surprising that we can draw the spectrum from $\omega = -\pi$ to $\omega = +\pi$ or equivalently from $f = -0.5$ to $f = 0.5$.
($\Rightarrow$ the frequency has a maximum in discrete time)

PREDICTION

We consider now the **prediction** problem for a **stationary process**.
(Kolmogorov - Wiener theory)

We can represent a stationary problem in 3 ways:
- covariance fun. $\gamma(\tau)$
- spectrum $\Gamma(\omega)$
- dynamic repr.:

$$\eta \sim WN \longrightarrow \boxed{W(z)} \xrightarrow{v}$$

For the prediction problem the best is the dynamical representation (it links the future with the past).

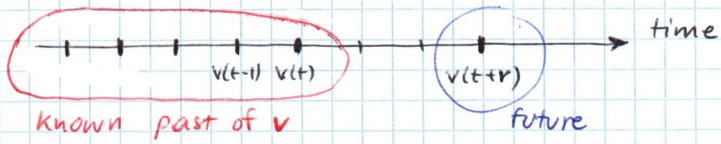
Consider a stationary process:

$$\eta \sim WN \longrightarrow \boxed{W(z)} \xrightarrow{v}$$

$$W(z) = w_0 + w_1 z^{-1} + w_2 z^{-2} + \dots \implies v(t) = w_0 \eta(t) + w_1 \eta(t-1) + w_2 \eta(t-2) + \dots$$

MA(∞) expression of the process $v(t)$

Problem: we want to find a fair estimate of $v(t+r)$ knowing the past of $v(\cdot)$ up to the time t: $v(t), v(t-1), v(t-2), \dots$



We face this problem in two steps:

1. **FAKE** prediction problem
   We assume that we know the past of η $\implies$ $\hat{v}(t+r|t)$
   prediction at time $t+r$ given observations up to time t (in this case observation of η)

2. **REAL** prediction problem
   We know the observations of the past of v $\implies$ $\hat{v}(t+r|t)$

---

1. FAKE PROBLEM

We can write $v(t+r)$ by the MA(∞) expression:

$$v(t+r) = \underbrace{w_0 \eta(t-r) + \dots + w_{r-1} \eta(t+1)}_{:= \alpha(t)} + \underbrace{w_r \eta(t) + w_{r+1} \eta(t-1) + \dots}_{:= \beta(t)} \dots$$

$$= \alpha(t) + \beta(t)$$

linear comb. of $\eta(t+1), \eta(t+2), \dots, \eta(t+r)$; it's the **FUTURE** of $\eta(\cdot)$ with respect to t

linear combination of $\eta(t-1), \eta(t-2), \dots$; it's the **PAST**

Observations:

- $\alpha(t) \perp \beta(t)$ ($\perp \equiv$ they're UNCORRELATED)
- Knowing the past of $\eta(\cdot)$ up to time t:
  - $\beta(t)$ can be computed
  - $\alpha(t)$ is fully unpredictable (the past of η doesn't bring any information for the future of η)

  In this case the only reasonable prediction of α is: $\hat{\alpha} = \mathbb{E}[\alpha] = 0$ (mean value)

$$\implies \hat{v}(t+r|t) = \beta(t) = w_r \eta(t) + w_{r-1} \eta(t-1) + \dots$$

- **PREDICTION ERROR**:

$$v(t+r) - \hat{v}(t+r|t) = \alpha(t) = \underbrace{w_0 \eta(t+r) + w_1 \eta(t+r-1) + \dots + w_{r-1} \eta(t+1)}_{\text{MA model: linear combination of } \eta(t+1), \eta(t+2), \dots, \eta(t+r)}$$

$\Longrightarrow$ since it's MA : $\text{Variance(error)} = (w_0^2 + w_1^2 + \dots + w_{r-1}^2)\,\lambda^2$

In particular :

- $r = 1$: $\text{Variance}_1(\text{error}) = w_0^2 \lambda^2$
- $r = 2$: $\text{Variance}_2(\text{error}) = (w_0^2 + w_1^2)\lambda^2$ $\quad (\geq \text{Var}_1(\text{error}))$
- $r = 3$: $\text{Var}_3(\text{error}) = (w_0^2 + w_1^2 + w_3^2)\lambda^2$ $\quad (\geq \text{Var}_2(\text{error}) \geq \text{Var}_1(\text{error}))$

$\Longrightarrow$ the variance of the error is monotonically increasing with the prediction horizon (if we want to predict a large number of steps ahead, then the prediction will be more uncert than the prediction of just one step ahead)

Finding the predictor in practice

$$\hat{v}(t+r|t) = w_r\,\eta(t) + w_{r+1}\,\eta(t-1) + \dots$$
$$= \underbrace{(w_r + w_{r+1} z^{-1} + \dots)}_{:= \hat{W}_r(z)}\,\eta(t)$$

$$\eta \sim WN \longrightarrow \boxed{\hat{W}_r(z)} \xrightarrow{\hat{v}(t+r|t)}$$

transfer function of the r-steps ahead predictor (**optimal** predictor)

$\Longrightarrow$ the problem is : given a dynamical model $\left(\xrightarrow{\eta} \boxed{W(z)} \xrightarrow{v}\right)$, i.e. given the transfer function $W(z)$, find the transfer function of the r-steps ahead predictor ($\hat{W}_r(z)$)

How can we procede ?
We consider the representation of $W(z)$ in negative powers of z :

$$W(z) = w_0 + w_1 z^{-1} + \dots + w_{r-1} z^{-r+1} + \underbrace{\boxed{w_r z^{-r} + w_{r+1} z^{-r-1} + \dots}}$$
$$= z^{-r}(w_r + w_{r+1} z^{-1} + \dots)$$
$$= z^{-r}\,\hat{W}_r(z)$$

$\Longrightarrow$ to find $\hat{W}_r(z)$ it is sufficient to do the long division

Example : AR(1) : $v(t) = a\,v(t-1) + \eta(t) \Longrightarrow W(z) = \dfrac{1}{1 - a z^{-1}} = \dfrac{z}{z-a}$

- 1-step predictor :

$$\begin{array}{l|l} z & z-a \\ \hline z-a & 1 \\ \hline / \; a & \end{array}$$

$$\Longrightarrow W(z) = 1 + \frac{a}{z-a} \cdot (z \cdot z^{-1}) = 1 + z^{-1} \underbrace{\frac{az}{z-a}}_{\hat{W}_1(z)}$$

- 2-step predictor :

$$\begin{array}{l|l} z & z-a \\ \hline z-a & 1 + a z^{-1} \\ \hline / \; a & \\ \quad a - a^2 z^{-1} & \\ \hline \quad / \; a^2 z^{-1} & \end{array}$$

$$\Longrightarrow W(z) = 1 + a z^{-1} + \frac{a^2 z^{-1}}{z-a} \cdot (z^2 \cdot z^{-2})$$
$$= 1 + a z^{-1} + z^{-2} \underbrace{\frac{a^2 z}{z-a}}_{\hat{W}_2(z)}$$

Remarks on the example:

- if we write $W(z) = \frac{C(z)}{A(z)}$ we can see that:

$$\hat{W}_1(z) = \frac{az}{z-a} = \frac{az}{A(z)} \leftarrow$$

$$\hat{W}_2(z) = \frac{a^2 z}{z-a} = \frac{a^2 z}{A(z)} \leftarrow$$

...

- We can write the predictor in the time domain:

1-step ahead: $\hat{W}_1(z) = \frac{az}{z-a} = \frac{a}{1-az^{-1}}$: $\xrightarrow{\eta} \boxed{\hat{W}_1(z)} \xrightarrow{\hat{v}(t+1|t)}$

$$\hat{v}(t+1|t) = \hat{W}_1(z)\,\eta(t) = \frac{a}{1-az^{-1}}\,\eta(t)$$

$$\Longrightarrow (1-az^{-1})\,\hat{v}(t+1|t) = a\,\eta(t)$$

$$\Longrightarrow \underbrace{\hat{v}(t+1|t)}_{\text{time domain expression of 1-step predictor}} = a\,\hat{v}(t|\underline{t-1}) + a\,\eta(t)$$

$$z^{-k}\,\hat{v}(t+r|t) = \hat{v}(t+r-k|t-k)$$

2-step ahead: $\hat{v}(t+2|t) = a\,\hat{v}(t+1|t-1) + a^2\,\eta(t)$

---

2. REAL PROBLEM

Problem: find the predictor from **data**, given the past of $v(\cdot)$.

How do we do it? We know how to determine the predictor knowing the past of $\eta(\cdot)$ ($\Rightarrow \hat{W}_r(z)$), but we don't know it, we know the past of $v(\cdot)$.

$\Longrightarrow$ problem: find the past values of $\eta(\cdot)$ knowing the past values of $v(\cdot)$:

$$\xrightarrow{\eta} \boxed{W(z)} \xrightarrow{v} \quad \Longrightarrow \quad \xrightarrow{v} \boxed{W^{-1}(z)} \longrightarrow ?$$

Can we invert the transfer function? Not always, it depends. (Delicate issue)

Example: $W(z) = \frac{3}{z-0.5} \Longrightarrow W^{-1}(z) = \frac{z-0.5}{3}$

$$\xrightarrow{u} \boxed{W^{-1}(z)} \xrightarrow{y} : \quad y(t) = \frac{z-0.5}{3}\,u(t)$$

$$\Longrightarrow \underbrace{y(t) = \frac{1}{3}u(t+1) - \frac{0{,}5}{3}u(t)}$$

the output depends on the future of the input
$\rightarrow$ this is **NOT** a dynamical system.
if we have a block: $\xrightarrow{I} \boxed{} \xrightarrow{O}$ means that the output is a function of the past input

Observations:

- main condition of invertibility: numerator and denominator of $W(z)$ must have the same degree
- main condition of stability: poles and zeros of $W(z)$ must be inside the unit disk (*)

(*) This is because if $W(z) = \frac{C(z)}{A(z)} \Longrightarrow W^{-1}(z) = \frac{A(z)}{C(z)}$: poles and zeros are the same of $W(z)$ but inverted (poles $(W(z))$ = zeros $(W^{-1}(z))$, zeros $(W(z))$ = poles $(W^{-1}(z))$, and we always want the poles inside the unit disk.

$\Longrightarrow$ We always consider the canonical spectral factor $\hat{W}(z) = \frac{\hat{C}(z)}{\hat{A}(z)}$. Properties :

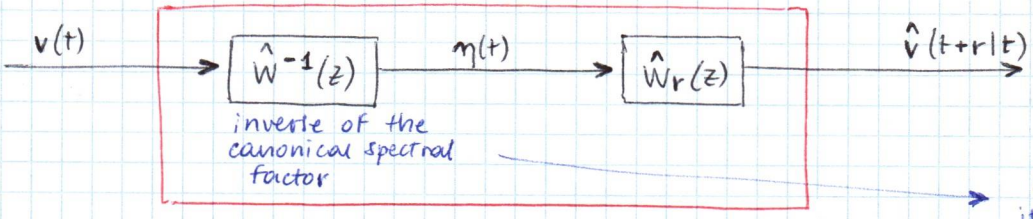
- degree of N and D are the same
  ($\Longrightarrow$ it's invertible)
- poles and zeros are inside the unit disk
  ($\Longrightarrow$ the inverse is stable)
- $\hat{A}$ and $\hat{C}$ are monic
- there are <u>no</u> common factor between $\hat{A}$ and $\hat{C}$

- $\hat{W}(z) = \frac{\hat{C}(z)}{\hat{A}(z)}$
- long division :
  $\hat{W}_r(z) = \frac{G_r(z)}{\hat{A}(z)}$ (from noise)

$$\boxed{W_r(z) = \hat{W}(z)^{-1}\, \hat{W}_r(z)}$$

predictor from DATA

Solution of the real prediction problem



inverse of the canonical spectral factor

WITHENING FILTER : it provides the WN from data

$:= W_r(z)$

transfer function of the optimal r-steps ahead predictor from **data**

$$\Longrightarrow \quad W_r(z) = \hat{W}^{-1}(z)\, \hat{W}_r(z) \qquad (*)$$

PREDICTOR
r STEPS AHEAD

it's used the solution of the fake prediction problem

Example : AR(1) : $W(z) = \frac{z}{z-a} = 1 + z^{-1}\frac{az}{z-a}$ ($|a|<1$, so the model is stable)

$v(t) = a\,v(t-1) + \eta(t)$

- poles and zeros $\in$ unit disc
- N & D same degree
- no common factor
- N & D are monic

1- step ahead predictor from **noise** η :

$$\hat{W}_1(z) = \frac{az}{z-a}$$

1- step ahead predictor from **data** **v** :

$$W_1(z) = \hat{W}^{-1}(z)\, \hat{W}_1(z) = \frac{\cancel{z-a}}{\cancel{z}} \cdot \frac{a\cancel{z}}{\cancel{z-a}} = a$$

$\Longrightarrow$ the optimal predictor from data **v** :

$$\hat{v}(t+1|t) = W_1(z)\, v(t) = a\,v(t)$$

$$\hat{v}(t+2|t) = a^2 v(t) \longrightarrow v(t+2) = a\,v(t+1) + \xi(t+2)$$

$$\hat{v}(t+k|t) = a^k v(t) \qquad = a(a\,v(t) + \xi(t+1)) + \xi(t+2)$$

$$= \underbrace{a^2 v(t)}_{\text{PAST}} + \underbrace{a\,\xi(t+1) + \xi(t+2)}_{\text{FUTURE}}$$

$$\Longrightarrow \hat{v}(t+2|t) = a^2 v(t)$$

$$(*) \quad W_r(z) = \frac{\cancel{A(z)}}{C(z)} \cdot \underbrace{\frac{G_r(z)}{\cancel{A(z)}}}_{\hat{W}_r(z)} = \frac{G_r(z)}{C(z)}$$

: the predictor from data can be easily obtained from the predictor from WN by replacing the denominator $A(z)$ with the numerator $C(z)$ of the canonical representation

Example: MA (1): $v(t) = \overset{1}{\cancel{c_0}}\,\eta(t) + c_1\,\eta(t-1)$

$$W(z) = \underset{1}{\cancel{c_0}} + c_1 z^{-1} = \frac{\overset{1}{\cancel{c_0}} z + c_1}{z}$$

$c_0 = 1$ to be monic

poles : $z = 0$
zeros : $z = -c$ $\Longrightarrow$ $|c| < 1$

$\Longrightarrow$ Canonical factor:

$$v(t) = \xi(t) + c\,\xi(t-1)$$

$$W(z) = \frac{z+c}{z}$$

1 - step ahead predictor : (from noise)

$$W(z) = w_0 + z^{-1}\hat{W}_1(z)$$

$$W(z) = 1 + z^{-1} c$$

$$\Longrightarrow \quad \hat{W}_r(z) = c$$

1 - step ahead predictor : (from data)

$$W_1(z) = \frac{c}{1+cz^{-1}}$$

in time domain:

$$v(t) \longrightarrow \boxed{W_1(z)} \xrightarrow{\hat{v}(t+1|t)}$$

$$\hat{v}(t+1|t) = \frac{c}{1+cz^{-1}}\, v(t)$$

$$\Longrightarrow (1+cz^{-1})\,\hat{v}(t+1|t) = c\,v(t)$$

$$\Longrightarrow \hat{v}(t+1|t) = -c\,\hat{v}(t|t-1) + c\,v(t)$$

2 - steps ahead :

$$\hat{v}(t+2|t) = 0 \quad \Longleftarrow \quad \hat{W}_2(z) = 0 \qquad (W(z) = 1 + cz^{-1} + z^{-2}\cdot 0)$$

Predictor of any ARMA process

$$v(t) = \underbrace{a_1 v(t-1) + \dots + a_{n_a} v(t-n_a)}_{AR(n_a)} + \underbrace{c_0\,\eta(t) + c_1\,\eta(t-1) + \dots + c_{n_c}\,\eta(t-n_c)}_{MA(n_c)}$$

$$\Longrightarrow \begin{cases} A(z) = 1 - a_1 z^{-1} - \dots - a_{n_a} z^{-n_a} \\ C(z) = c_0 + c_1 z^{-1} + \dots + c_{n_c} z^{-n_c} \end{cases}$$

$$\Longrightarrow \quad W(z) = \frac{C(z)}{A(z)} = \frac{\overset{1}{c_0} z^n + c_1 z^{n-1} + \dots}{z^n - a_1 z^{n-1} - \dots}$$

we multiply by $n = \max(n_a, n_c)$

Is this form a canonical representation? Yes, if $c_0 = 1$.
so we assume that $W(z)$ is canonical

$$W(z) \longrightarrow \hat{W}(z)$$

To find the predictor we use the long division. How many steps? As many as the prediction horizon we want to compute.

1 - step ahead:

$$\begin{array}{l|l} C(z) & A(z) \\ \underline{A(z)} & \overline{1} \\ C(z)-A(z) & \end{array}$$

because both are monic

Hence:

$$W(z) = 1 + \frac{C(z)-A(z)}{A(z)} = 1 + z^{-1}\,\frac{z(C(z)-A(z))}{A(z)}$$

$\Longrightarrow$ Transfer function of the predictor from noise (ξ):

$$\hat{W}_1(z) = \frac{z(C(z)-A(z))}{A(z)}$$

denominator of the canonical spectral factor

$\Longrightarrow$ Transfer function of the predictor from data (v)

$$W_1(z) = \frac{z(C(z)-A(z))}{C(z)}$$

numerator of the canonical spectral factor

In time domain:

$$v(t) \longrightarrow \boxed{W_1(z)} \longrightarrow \hat{v}(t+1|t)$$

$$\hat{v}(t+1|t) = \frac{z(C(z)-A(z))}{C(z)}\,v(t) \Longrightarrow C(z)\,\hat{v}(t+1|t) = (C(z)-A(z))\,v(t+1)$$

!

How can it be $v(t+1)$?
We use the data at time $t+1$ to make prediction of the data at time $t+1$? **NO.**

$$C(z)-A(z) = \cancel{1} + c_1 z^{-1} + \dots + \\ -(\cancel{1} - a_1 z^{-1} - \dots)$$

$$\Longrightarrow \hat{v}(t+1|t) = -c_1\,\hat{v}(t|t-1) - \dots + \\ + (a_1+c_1)\,v(t-1) + (a_2+c_2)\,v(t-2) + \dots$$

Alternatively (we want the same result more easily):

Shortcut derivation of optimal predictor

ARMA process model: $A(z)\,v(t) = C(z)\,\xi(t)$

$$\pm\,\hat{C}(z)\,v(t)$$

$$\Longrightarrow C(z)\,v(t) = (C(z)-A(z))\,v(t) + C(z)\,\xi(t)$$

$$\Longrightarrow v(t) = \frac{C(z)-A(z)}{C(z)}\,v(t) + C(z)\,\xi(t)$$

this is a function of only the past
since both $C(z)$ and $A(z)$ are monic
$\Longrightarrow$ this is a function of $v(t-1), v(t-2), \dots$

$$\Longrightarrow \hat{v}(t|t-1) = \frac{C(z)-A(z)}{C(z)}\,v(t)$$

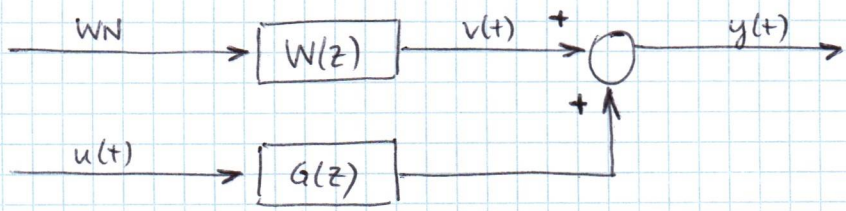
optimal **1** step ahead predictor for an ARMA model

We can also write it as: $C(z)\,\hat{v}(t|t-1) = (C(z)-A(z))\,v(t)$

EXOGENOUS VARIABLES
(From ARMA to ARMAX)

Up to now we considered stochastic stationary processes $v(t)$: $\xrightarrow{WN} \boxed{W(z)} \xrightarrow{v(t)}$

We want now to generalize this situation to the case of some external variables influencing the signal v (external variables / exogenous variables)



The signal $y(t)$ is determined by a stochastic process plus a signal produced by an external (exogenous) variable.

Example: **AR(1)** : $v(t) = a\, v(t-1) + \xi(t)$

Optimal predictor : $\hat{v}(t+1|t) = a\, v(t)$

ARX : $v(t) = a\, v(t-1) + u(t) + \xi(t)$

Optimal predictor ?

Assume that $u(t) = \text{constant} = \bar{u}$

Compute : $\mathbb{E}[v(t)] = m$:

$$m = \mathbb{E}[a\, v(t-1)] + \bar{u} + 0 = a\, \mathbb{E}[v(t-1)] + \bar{u}$$

$$\Longrightarrow (1-a)\, m = \bar{u}$$

$$\Longrightarrow m = \frac{\bar{u}}{1-a}$$

Now we define: $\tilde{v}(t) = v(t) - m$, equation for $\tilde{v}$?

$$v(t) = a\, v(t-1) + \bar{u} + WN$$

$$\Longrightarrow \tilde{v}(t) + m = a(\tilde{v}(t-1) + m) + \bar{u} + WN$$

$$\Longrightarrow \tilde{v}(t) = a\, \tilde{v}(t-1) + WN$$

$\Longrightarrow$ $\tilde{v}(\cdot)$ satisfies a traditional AR(1)

$\Longrightarrow$ optimal predictor (for $\tilde{v}$) : $\hat{\tilde{v}}(t+1|t) = a\, \tilde{v}(t)$

We want an optimal predictor for v :

$$\hat{v}(t+1|t) = \hat{\tilde{v}}(t+1|t) + m$$

$$= a\, v(t) - am + m$$

$$= a\, v(t) + \bar{u}$$

(ARX) optimal predictor : $\hat{v}(t+1|t) = a\, v(t) + \bar{u}$

AR(1) : $v(t+1) = a\, (t) + WN \Longrightarrow \hat{v}(t+1|t) = a\, v(t)$

ARX : $v(t+1) = a\, v(t) + \bar{u} + WN \Longrightarrow \hat{v}(t+1|t) = a\, v(t) + \bar{u}$

There is also: ARMA $\rightarrow$ ARMAX :

ARMA : $A(z)\,v(t) = C(z)\,\xi(t)$

ARMAX : $A(z)\,v(t) = C(z)\,\xi(t) + B(z)\,u(t)$

MODELS

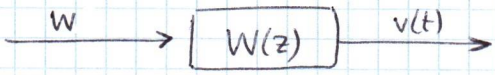
We have two type of models:

(a.) models for time series: unique signal $v(t)$

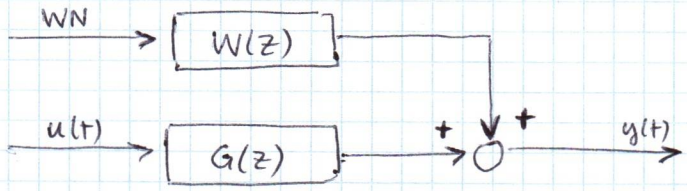
(b.) models for systems: input signal $u(t)$, output signal $y(t)$

We want to find a model s.t. :

a.



b.



"all models are wrong, some are useful"

BOX & JENKINS model (most general linear description of signal y including uncertainty)

Special cases:

- AR MODEL (for time series)

$$v(t) = a_1 v(t-1) + \dots + a_{n_a} v(t-n_a) + \xi(t)$$

$$W(z) = (1 - a_1 z^{-1} - \dots - a_{n_a} z^{-n_a})^{-1} = \frac{1}{A(z)}$$

- MA MODEL

$$v(t) = \xi(t) + c_1 \xi(t-1) + \dots + c_{n_c} \xi(t-n_c) \qquad (c_0 = 1)$$

$$W(z) = 1 + c_1 z^{-1} + \dots + c_{n_c} z^{-n_c} = C(z)$$

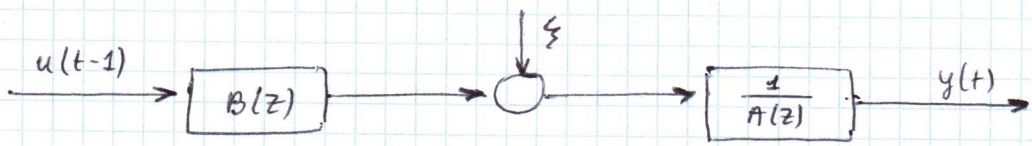
- ARX MODEL

$$y(t) = a_1 y(t-1) + \dots + a_{n_a} y(t-n_a) + b_1 u(t-1) + \dots + b_{n_b} u(t-n_b) + \xi(t)$$

$$A(z) = a_1 z^{-1} + \dots + a_{n_a} z^{-n_a}$$

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

$$A(z)\,y(t) = B(z)\,u(t-1) + \xi(t)$$



$$W(z) = \frac{1}{A(z)} \qquad \text{(from } \xi(t) \text{ to } y(t)\text{)}$$

$$G(z) = \frac{B(z)}{A(z)} \qquad \text{(from } u(t-1) \text{ to } y(t)\text{)}$$

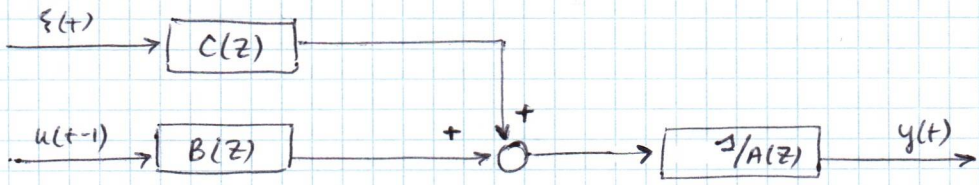
- ARMA MODEL

$$v(t) = a_1 v(t-1) + \dots + a_{n_a} v(t-n_a) + (t) + c_1 \xi(t-1) + \dots + c_{n_c} \xi(t-n_c)$$

$$W(z) = \frac{C(z)}{A(z)}$$

- **ARMAX MODEL**

$$A(z)\,y(t) = C(z)\,\xi(t) + B(z)\,u(t-1)$$



$$W(z) = \frac{C(z)}{A(z)}$$

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

PREDICTOR (Box & Jenkins)

we consider the B & J model:

$$y(t) = \underbrace{G(z)\,u(t-1)}_{\text{deterministic term}} + \underbrace{W(z)\,\xi(t)}_{\text{stochastic term}}$$

$$\xrightarrow{\cdot\left(\frac{1}{W(z)}\right),\ \pm y(t)} \quad \frac{1}{W(z)}\,y(t) \pm y(t) = \frac{G(z)}{W(z)}\,u(t-1) + \xi(t)$$

$$\Longrightarrow \quad y(t) = \overset{1}{\boxed{\left(1-\frac{1}{W(z)}\right)y(t)}} + \overset{2}{\boxed{\frac{G(z)}{W(z)}\,u(t-1)}} + \overset{3}{\boxed{\xi(t)}}$$

If $W(z)$ is the canonical factor then N & D of $W(z)$ are monic polynomial of the same degree:

$$\frac{1}{W(z)} = 1 + \alpha_1 z^{-1} + \alpha_2 z^{-2} + \dots$$

since $\frac{1}{W(z)}$ is the quotient of two monic polynomials with the same degre.

$$\Longrightarrow \left(1-\frac{1}{W(z)}\right) = \cancel{1} - \cancel{1} - \alpha_1 z^{-1} - \alpha_2 z^{-2} \dots$$

$$\Longrightarrow \left(1-\frac{1}{W(z)}\right) y(t) = f\left(y(t-1), y(t-2), \dots\right)$$

so **1** depends only on the <u>past</u>

Function of the past of $u(\cdot)$

sample of WN at time t (fully unpredictable from the past)

1 + **2** are perfectly predictable and computable.
3 is completely unpredictable and **1** from the past.

$\Longrightarrow$ best predictor: **1 + 2**

<u>B & J model predictor</u>:

$$\boxed{\hat{y}(t|t-1) = \left(1-\frac{1}{W(z)}\right)y(t) + \frac{G(z)}{W(z)}\,u(t-1)}$$

from this expression we can derive the predictor of all the models we have introduced

Example: ARX model: $W(z) = \frac{1}{A(z)}$, $G(z) = \frac{B(z)}{A(z)}$

$$\hat{y}(t|t-1) = (1-A(z))\, y(t) + \frac{B(z)}{\cancel{A(z)}} \cancel{A(z)}\, u(t-1)$$

$$\longrightarrow \hat{y}(t|t-1) = (1-A(z))\, y(t) + B(z)\, u(t-1)$$

$$\longrightarrow \hat{y}(t|t-1) = a_1 y(t-1) + a_2 y(t-2) + \dots + a_{n_a} y(t-n_a) + b_1 y(t-1) + b_2 y(t-2) + \dots + b_{n_b} y(t-n_b)$$

So $\hat{y}(t|t-1)$ is linear in the parameters.

$$\underbrace{\phi(t)}_{\text{vector of observations}} = [\, y(t-1), \dots, y(t-n_a),\ u(t-1), \dots, u(t-n_b)\,]^T$$

$$\underbrace{\theta}_{\text{vector of parameters}} = [\, a_1, \dots, a_{n_a}, b_1, \dots, b_{n_b}\,]^T$$

$$\longrightarrow \hat{y}(t|t-1) = \phi(t)^T \theta = \theta^T \phi(t)$$

Example: ARMAX model: $W(z) = \frac{C(z)}{A(z)}$, $G(z) = \frac{B(z)}{A(z)}$

$$\hat{y}(t|t-1) = \left(1 - \frac{A(z)}{C(z)}\right) y(t) + \frac{B(z)}{\cancel{A(z)}} \cdot \frac{\cancel{A(z)}}{C(z)}\, u(t-1)$$

$$\Longrightarrow \hat{y}(t|t-1) = \left(1 - \frac{A(z)}{C(z)}\right) y(t) + \frac{B(z)}{C(z)}\, u(t-1)$$

$$= \frac{C(z) - A(z)}{C(z)}\, y(t) + \frac{B(z)}{C(z)}\, u(t-1)$$

$$\phi(t) = [\, y(t-1), \dots, y(t-n_a),\ u(t-1), \dots, u(t-n_b)\,]^T$$

$$\theta = \left[\, \frac{c_1 - a_1}{c_1}, \dots, c_n \right.$$

problem:
we have to show that it's no longer linear in the parameters

PART II : IDENTIFICATION

Problem : use the data to construct the model.
There are two type of problems :

- physical models (based on scientific laws)
- black box model (based on correlation approach)

We focus on the black box model : find the relationships among variables without knowing the internal system.

There are two ways to procede :

1. data $\longrightarrow$ estimate of covariance function $\gamma(\tau)$ or spectrum $\Gamma(\omega)$ $\rightarrow$ model
2. direct estimation of the model's parameters

we focus on this

BASIC IDEA : use predictive identification, i.e. a model in a given class is good in extent to which the prediction error is <u>nice</u>

"Nice" ?
In the previous chapter a "nice" property of the prediction error was the whiteness.
We now relax a little and we look for a "small" prediction error.

The basic performance index will be $\varepsilon_\theta(t)$:

$\varepsilon_\theta(t)$ = prediction error at time t of a certain model $M(\theta)$
(where θ is the vector of the model parameters)

What we are trying to minimize is the average prediction error squared :

$\Longrightarrow$ <u>problem</u> :

$$\begin{cases} \min\limits_{\theta} J = \min\limits_{\theta} \frac{1}{N} \sum \varepsilon_\theta(t)^2 \\ \hat{\theta} = \arg\min\limits_{\theta} J \end{cases}$$

LEAST SQUARES IDENTIFICATION METHOD (LS)

Data $\longrightarrow$ family of models $\longrightarrow$ criterion of fit

We have [1] <u>data</u> : (observations /) measurements of past samples over a certain time interval (for example $t = 1, 2, \dots, N$).

We select a [2] <u>family of models</u> to fit data :

- AR model for time series
- ARX model for systems

We consider a [3] <u>criterion of fit</u> : $\min\limits_{\theta} J_N(\theta) = \min\limits_{\theta} \frac{1}{N} \sum\limits_{t=1}^{N} \varepsilon_\theta(t)^2$

The problem is characterized by these 3 items.

Solving the LS problem

<u>ARX</u> : $M(\theta)$: $y(t) = a_1 y(t-1) + \dots + a_{n_a} y(t-n_a) +$
$+ b_1 u(t-1) + \dots + b_{n_b} u(t-n_b) + \xi(t)$

model associated to θ

ARX: θ parameter vector, $\phi(t)$ observation vector:

$$y(t) = \phi(t)^T \theta + \varepsilon(t) = \theta^T \phi(t) + \varepsilon(t)$$

We pass to the model in predictive form:

$$\hat{M}(\theta) = \hat{y}(t|t-1) = \phi(t)^T \theta = \theta^T \phi(t)$$

with the criterion:

$$\min_\theta J(\theta) = \min_\theta \frac{1}{N} \sum_{t=1}^N \varepsilon_\theta(t)^2 \qquad (\varepsilon_\theta(t) = y(t) - \phi(t)^T \theta)$$

$\Longrightarrow$ solution: $\frac{d}{d\theta} J(\theta) = \underline{0} \Longrightarrow \hat{\theta}$

the derivative of a scalar w.r.t. a vector is a vector

$$\frac{dJ}{d\theta} = \left[\frac{\partial J}{\partial a_1}, \frac{\partial J}{\partial a_2}, \dots, \frac{\partial J}{\partial b_1}, \frac{\partial J}{\partial b_2}, \dots \right]$$

we consider it as a **row** vector

$$\Longrightarrow \frac{d}{d\theta} \frac{1}{N} \sum_{t=1}^N (y(t) - \phi(t)^T \theta)^2 = \frac{2}{N} \sum_{t=1}^N (y(t) - \phi(t)^T \theta)(-\phi(t)^T)$$

Setting $= 0$:

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

$$\sum_{t=1}^N (y(t)\, \phi(t)^T)^T = \sum_{t=1}^N ((\phi(t)^T \theta)\, \phi(t)^T)^T$$

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

NORMAL EQUATIONS

computable from data — computable from data

linear system in the unknown θ: $\square\, \theta = \square$

$\sum_t \phi \phi^T$ $\quad$ $\sum \phi y$

<u>Question:</u> a solution $\hat{\theta}$ of the normal equation is a point of minimum?

Computing $\frac{d^2 J}{d\theta^2}$ we see that it's positive semidefinite $\left(\left[\frac{\partial^2 J}{\partial \theta_i \partial \theta_j} \right]_{ij} \right)$

(<u>Ex.</u>) $$\frac{d^2 J}{d\theta^2} = \frac{2}{N} \sum_{t=1}^N \phi(t)\, \phi(t)^T$$

(<u>Ex.</u>) $$\frac{d^2 J}{d\theta^2} \geq 0$$

<u>Conclusion</u>: the normal equation may have 1 or many solutions.

If the matrix $*$ is invertible, the $\exists!$ solution to the normal eq.:

$$\hat{\theta} = \left(\sum_t \phi(t)\, \phi(t)^T\right)^{-1} \sum_t \phi(t)\, y(t) .$$

If $*$ is not invertible ($\equiv$ singular) $\Longrightarrow$ the normal equation has ∞-many solutions. All such solutions are equivalent to each other in the sense that $J(\theta)$ takes the same minimal value for all such θ solving the normal equation.

Observation:

$$J(\theta) = \frac{1}{N} \sum_t \varepsilon_\theta(t)^2 = \frac{1}{N} \sum_t \left(y(t) - \phi(t)^T \theta\right)^2$$ quadratic in θ

Taylor series developement:

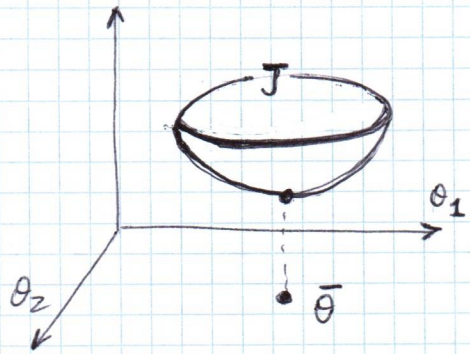
$$J(\theta) = J(\bar\theta) + (\theta - \bar\theta) \cancel{\frac{dJ(\theta)}{d\theta}\Big|_{\theta=\bar\theta}} + \frac{(\theta-\bar\theta)^2}{2} \frac{d^2 J(\theta)}{d\theta^2}\Big|_{\theta=\bar\theta} + \cancel{\dots}$$

$\bar\theta$ any solution of the normal equation

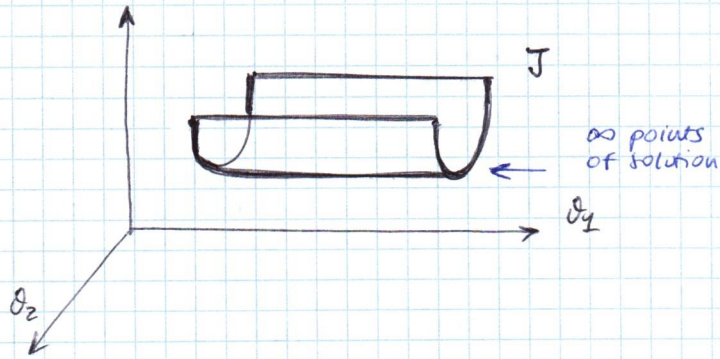
if $\bar\theta$ is a solution then $\frac{dJ}{d\theta}\Big|_{\theta=\bar\theta} = 0$ since the solutions are obtained by setting $= 0$ the 1st derivative

since J is quadratic in θ then there are no components of order > 2

$$\longrightarrow \quad J(\theta) = J(\bar\theta) + \frac{1}{2}(\theta - \bar\theta)^2 \frac{d^2 J}{d\theta^2}\Big|_{\theta=\bar\theta}$$



$$\boxed{\frac{d^2 J}{d\theta^2} > 0}$$



$$\boxed{\frac{d^2 J}{d\theta^2} \geq 0}$$

(RECAP) Problem of solving normal equations

ARX in predictive form: $\hat y(t) = a_1 y(t-1) + \dots + a_{n_a} y(t-n_a) +$ AR

$+ b_1 u(t-1) + \dots + b_{n_b} y(t-n_b)$ X

We define:

$$\theta = \begin{bmatrix} a_1 \\ \vdots \\ a_{n_a} \\ b_1 \\ \vdots \\ b_{n_b} \end{bmatrix}, \qquad \phi(t) = \begin{bmatrix} y(t-1) \\ \vdots \\ y(t-n_a) \\ u(t-1) \\ \vdots \\ u(t-n_b) \end{bmatrix}$$

parameters vector (unknown)

data: observation vector (known)

$$\hat\theta = \arg\min_\theta \frac{1}{N} \sum_{t=1}^{N} \varepsilon_\theta(t)^2 \quad \left(:= \arg\min_\theta J(\theta)\right)$$

N = number of data
θ = par. vector of model $M(\theta)$

Method to find $\hat\theta$: by solving the normal equations:

$$\underbrace{\left(\sum_{t=1}^{N} \phi(t)\, \phi(t)^T\right)}_{:= S(N)} \theta = \sum_{t=1}^{N} \phi(t)\, y(t)$$

$$\underbrace{\frac{1}{N}\left(\sum_{t=1}^{N} \phi(t)\,\phi(t)^T\right)}_{:= R(N) = \frac{1}{N} S(N)} \theta = \frac{1}{N} \sum_{t=1}^{N} \phi(t)\, y(t)$$

Solution : if $S(N)$ (or $R(N)$), which is a $n \times n$ matrix ($n = n_a + n_b$) is invertible then there is an unique solution given by :

$$\boxed{\hat{\theta} = \left(\sum_{t=1}^{N} \phi(t)\,\phi(t)^T\right)^{-1} \sum_{t=1}^{N} \phi(t)\, y(t)} \quad := \text{LS FORMULA}$$

If $S(N)$ (or $R(N)$) are singular, then there are ∞-many models with the same predictive capability (the data can be explained by many models.

PROBLEM OF IDENTIFIABILITY ($\equiv$ INVERTIBILITY OF $S(N)$ and $R(N)$)
(Corresponds to the uniqueness of the solution of normal equations)

Is the information provided by data sufficient to estimate all parameters?

Example : data generated by a system S : $\underset{\text{transfer function from } u \text{ to } y}{G^o(z)} = \dfrac{z}{(z+0.5)(z+0.8)}$

The problem is to identify parameters of $G^o(z)$ from data (the coefficients 1, 0.5, 0.8, ..)

Suppose we perform an identification experiment during which

$u(t) = \text{constant} = \bar{u}$:

$$\xrightarrow{u(t)} \boxed{G^o(z)} \xrightarrow{y(t)}$$

If u is constant then, being the system stable, in the long run y will be constant and related to the input as :

$$\bar{y} = \mu\, \bar{u}$$

$:=$ GAIN OF THE SYSTEM (S)

Recall : $\underset{\text{gain of the system}}{\mu} = G^o(z)\big|_{z=1}$ (see appendix in MIDA book)

We can estimate : $\mu = \dfrac{\bar{y}}{\bar{u}}$

In this experiment the input u is not rich enough to identify all the parameters. $\longrightarrow$ Data do not bring sufficient informations to obtain a unique estimation of θ.

PERSISTENT EXCITATION

$$R(N) = \frac{1}{N} \sum_{t=1}^{N} \phi(t)\, \phi(t)^T$$

We consider an ARX(1,1) model :

$$\phi(t) = \begin{bmatrix} y(t-1) \\ u(t-1) \end{bmatrix} \longrightarrow \phi(t)\,\phi(t)^T = \begin{bmatrix} y(t-1)^2 & \cdots \\ \cdots & u(t-1)^2 \end{bmatrix}$$

$$\Longrightarrow \quad R(N) = \frac{1}{N} \sum \phi(t)\,\phi(t)^T$$

$$= \begin{bmatrix} \frac{1}{N}\sum y(t-1)^2 & \cdots \\ \cdots & \frac{1}{N}\sum u(t-1)^2 \end{bmatrix}$$

sample variance of y

sample variance of u (we assume mean = 0) when $N \to \infty$: $\frac{1}{N}\sum u(t-1)^2 \to \gamma_u(0)$

$$\Longrightarrow \quad \bar{R} = \lim_{N\to\infty} R(N) = \begin{bmatrix} \gamma_y(0) & \cdots \\ \cdots & \gamma_u(0) \end{bmatrix}$$

We consider now a (generic) ARX(n_a, n_b) model :

$$\bar{R} = \begin{bmatrix} \bar{R}_y & \cdots \\ \cdots & \bar{R}_u \end{bmatrix} \in \mathbb{R}^{n \times n} \qquad (n = n_a + n_b)$$

$$\bar{R}_u = \begin{bmatrix} \gamma_u(0) & \gamma_u(1) & \gamma_u(2) & \cdots \\ \gamma_u(1) & \gamma_u(0) & \gamma_u(1) & \cdots \\ \gamma_u(2) & \gamma_u(1) & \gamma_u(0) & \cdots \\ \vdots & & & \end{bmatrix} =$$

TOEPLITZ MATRIX

(HANKEL MATRIX)

Observations :

- $\bar{R}_u$ depends only on $u(t)$
- The invertibility of $\bar{R}_u$ is necessary for the invertibility of $\bar{R}$

Def. a signal $u(\cdot)$ is PERSISTENTLY EXCITING of order k if the $k \times k$ matrix :

$$\bar{R}_u = \begin{bmatrix} \gamma_u(0) & \gamma_u(1) & \gamma_u(2) & \cdots \\ \gamma_u(1) & \gamma_u(0) & \gamma_u(1) & \cdots \\ \gamma_u(2) & \gamma_u(1) & \gamma_u(0) & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix} \quad \text{is invertible.}$$

$\Longrightarrow$ In order to obtain an unique estimate of vector θ, it is necessary that $u(\cdot)$ be persistently exciting of order n_b.

Example :

$$\xrightarrow{u(t)} \boxed{S} \xrightarrow{y(t)} \qquad u(t) \sim WN(0, \lambda^2)$$

Data $u, y \Longrightarrow \hat{\theta}$ (by solving normal equations)

$$\bar{R}_u = \begin{bmatrix} \lambda^2 & \cancel{\gamma_u(1)}\,0 & 0 & 0 & \cdots \\ 0 & \lambda^2 & 0 & 0 & \cdots \\ 0 & 0 & \lambda^2 & 0 & \cdots \\ \cdots & & & & \end{bmatrix} = \lambda^2 I$$

$\Longrightarrow$ **WN** is persistently exciting of any order $\Longrightarrow$ so it's a nice signal for identification because it brings all the infos in the data

UNIQUENESS OF ESTIMATION

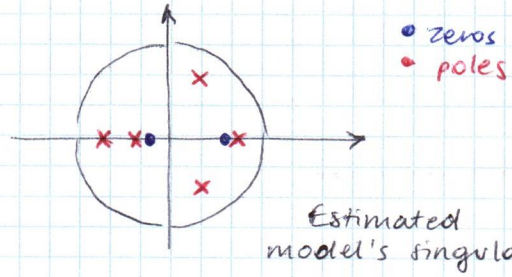
- $u(\cdot)$ must be rich enough (experimental identifiability)
- structural identifiability

Suppose we have a system S of order 2 and a model $M(\theta)$ of order 5.
Can we estimate θ in an unique way?
The model is oversized (too many parameters) $\longrightarrow$ it can be semplified.
Engineering approach:

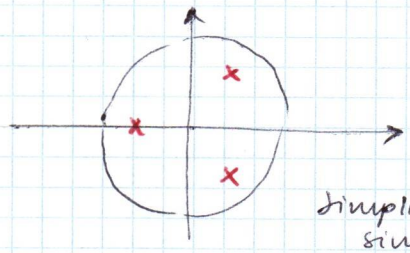
- identify $M(\theta)$ and consider poles and zeros of the estimated model:

  the zeros and poles that are really close to each other simplify themselves:

- simplified model:



Estimated model's singularities



Simplified model's singularities

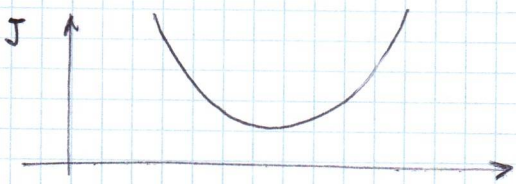
MAXIMUM LIKELIHOOD METHOD (ML)

(Correspondent of the LS method)

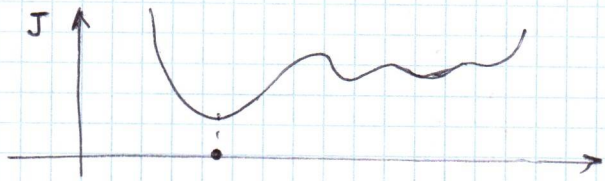
LS : AR, ARX : the model is linear in the parameters $\longrightarrow$ normal equations

ML : ARMA, ARMAX : the model is **non** linear in the parameters

Problem: $$\min_{\theta} J(\theta) = \min_{\theta} \frac{1}{N}\sum_{t=1}^{N} \underbrace{\varepsilon_\theta(t)}_{\text{prediction error of the model } M(\theta)}{}^2$$



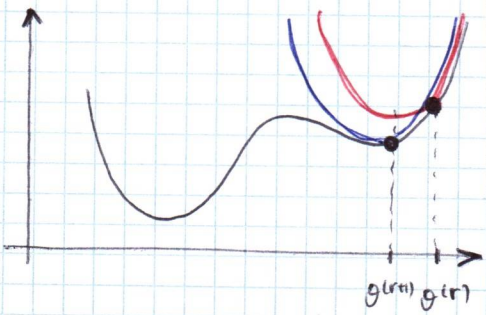
AR / ARX case



ARMA / ARMAX case

$\Longrightarrow$ we want to find the minimum of a non quadratic function

How? ITERATIVE NEWTON PROCEDURE



Suppose we start from r $(\theta^{(r)})$. We try to approximate the real function with the quadratic function and then we consider the minimum of this quadratic function and we use it as second iteration $(r+1 : \theta^{(r+1)})$.
And then we go on.

However, there is no guarantee to reach the global minimum. We can start from different points, reach a lot of local minima and then take the minimum among the minima. (Still no guarantee that we take the global one).

We denote this quadratic approximation function with $V(\theta)$.
How can we find proxy $V(\theta)$?

- $V(\theta)\big|_{\theta^{(r)}} = J(\theta)\big|_{\theta^{(r)}}$
- $\dfrac{dV(\theta)}{d\theta}\Big|_{\theta^{(r)}} = \dfrac{dJ(\theta)}{d\theta}\Big|_{\theta^{(r)}}$
- $\dfrac{d^2V(\theta)}{d\theta^2}\Big|_{\theta^{(r)}} = \dfrac{d^2J(\theta)}{d\theta^2}\Big|_{\theta^{(r)}}$

In this way $V(\theta)$ is completely defined and we can find $\theta^{(r+1)}$

→ Newton formula:

$$\begin{cases} \theta^{(r+1)} = \theta^{(r)} + \left(\sum_t \phi(t)\,\phi(t)^T\right)^{-1} \sum_t \psi(t)\,\varepsilon(t) \\ \psi(t) = -\dfrac{dJ(\theta)}{d\theta}\Big|_{\theta^{(r)}} \\ \varepsilon(t) := \text{prediction error evaluated with model } M(\theta)\big|_{\theta=\theta^{(r)}} \end{cases}$$

The issue is then how to compute $\psi(t)$ and $\varepsilon(t)$ from data by exploiting the model estimated with $\theta^{(r)}$.

PERFORMANCE OF IDENTIFICATION MODELS (ASYMPTOTIC ANALYSIS OF PEM METHODS)

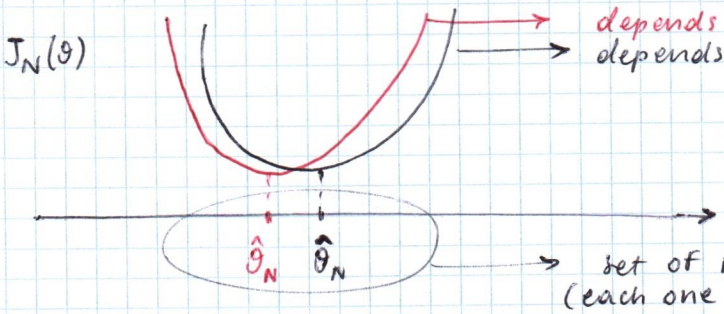
Summarizing PE identification method (general framework of PEM):

- data
- family of models
- predictive criterion : $\hat{\theta} = \arg\min_{\theta} J_N(\theta) = \arg\min_{\theta} \frac{1}{N}\sum_{t=1}^{N} \varepsilon_\theta(t)^2$

The optimal model is $M(\hat{\theta})$.

$\varepsilon_\theta(t) = y(t) - \hat{y}_\theta(t)$

How optimal is the optimal model?

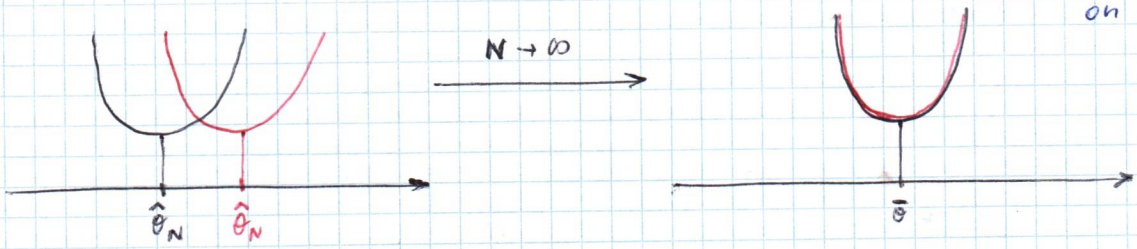


→ depends on data
→ depends on date

→ set of minima (each one is associated with a set of data)

data are provided by an outcome of some random experiment (s)

$N \to \infty \Longrightarrow J_N(\theta) \longrightarrow \bar{J} = \mathbb{E}_s[\varepsilon_\theta(t)^2]$ = average over all possible s (it doesn't depend on s anymore)



$J_N(\theta) \xrightarrow{N\to\infty} \bar{J}(\theta)$

$\hat{\theta} \xrightarrow{N\to\infty} \bar{\theta}$

MAXIMUM LIKELIHOOD IDENTIFICATION METHOD

Problem of estimating the parameters of ARMA/ARMAX models
(LS to estimate AR/ARX models)

Consider the system:

$$A(z)\,y(t) = B(z)\,u(t-1) + C(z)\,\eta(t)$$

$$A(z) = 1 - a z^{-1}$$
$$B(z) = b$$
$$C(z) = 1 + c z^{-1}$$

$\Longrightarrow$ predictor: $C\hat{y} = (C-A)\,y + B\,u$

$\Longrightarrow$ transfer function from y to $\hat{y}$:

$$\frac{C-A}{C} = \frac{a+c}{1+cz^{-1}} = (a+c) - c(a+c)z^{-1} - \dots$$

long division

$$\Longrightarrow \hat{y}(t) = (a+c)\,y(t-1) - c(a+c)\,y(t-2) - \dots$$

no more linear in the parameters

Minimization of $J = \frac{1}{N}\sum_{t=1}^{N} \varepsilon(t)^2$?

$M(\theta)$ model: $A(z)\,y(t) = B(z)\,u(t-1) + C(z)\,\eta(t)$ — model

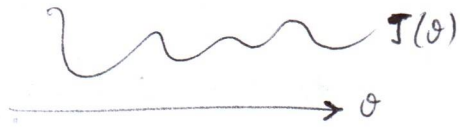
$\hat{M}(\theta)$ model in predictive form: $C(z)\,\hat{y}(t) = (C(z) - A(z))\,y(t) + B(z)\,u(t-1)$ — predictor

$$\Longrightarrow \underset{\substack{\text{vector of}\\ \text{parameters}\\ \text{(to be estimated)}}}{\theta} = \begin{bmatrix} a_1 \\ \vdots \\ a_{n_a} \\ b_1 \\ \vdots \\ b_{n_b} \\ c_1 \\ \vdots \\ c_{n_c} \end{bmatrix} \begin{matrix} \left.\right\} A(z) \\ \left.\right\} B(z) \\ \left.\right\} C(z) \end{matrix}$$

We want to estimate θ from data $\Rightarrow$ we use an iterative minimization procedure ($\min J(\theta)$)

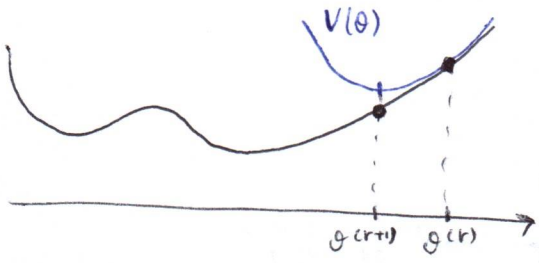
note: in LS

, in ML



(since it is <u>not</u> linear in params.)

NEWTON ITERATIVE METHOD:



$V(\theta)$: quadratic approximation function

$$\begin{cases} V(\theta)\big|_{\theta(r)} = J(\theta)\big|_{\theta(r)} \\ \text{grad}(V(\theta))\big|_{\theta(r)} = \text{grad}(J(\theta))\big|_{\theta(r)} \\ \text{hessian}(V(\theta))\big|_{\theta(r)} = \text{hessian}(J(\theta))\big|_{\theta(r)} \end{cases}$$

Paraboloid minimum:

$$\vartheta^{(r+1)} = \vartheta^{(r)} - \left(\frac{d^2 J}{d\vartheta^2}\Big|_{\vartheta^{(r)}} \right)^{-1} \left(\frac{\partial J}{\partial \vartheta}\Big|_{\vartheta^{(r)}} \right)^T$$

$$\| = \| - \square \cdot (\text{—})^T$$

We need to compute <u>gradient</u> and <u>hessian</u>:

$$J(\vartheta) = \frac{1}{N} \sum \varepsilon(t)^2$$

$$\frac{dJ(\vartheta)}{d\vartheta} = -\frac{2}{N} \sum \varepsilon(t)\, \psi(t)$$

$$\frac{d^2 J(\vartheta)}{d\vartheta^2} = \frac{2}{N} \sum \psi(t)\, \psi(t)^T - \frac{2}{N} \sum \varepsilon(t) \frac{d^2 \varepsilon(t)}{d\vartheta^2}$$

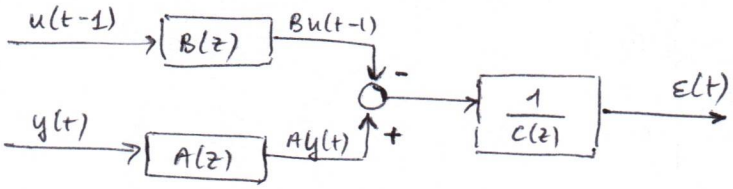
$$\begin{cases} \varepsilon(t) = y(t) - \hat{y}(t) \\ \psi(t) = -\left(\dfrac{d\varepsilon(t)}{d\vartheta} \right)^T \end{cases}$$

this term is usually NEGLECTED (for numerical complications, moreover, by eliminating this, we guarantee that $\frac{d^2 J}{d\vartheta^2} \geq 0$)

How to compute $\varepsilon(\cdot)$ and $\psi(\cdot)$ from data?
GAUSS-NEWTON ALGORITHM (which neglects)

$\varepsilon(t)$: (from data) Optimal predictor: $C\hat{y}(t) = (C-A)\,y(t) + B\,u(t-1)$

$$\Longrightarrow \boxed{C\,\varepsilon(t) = A\,y(t) - B\,u(t-1)}_{(*)} \Longrightarrow \varepsilon(t)$$



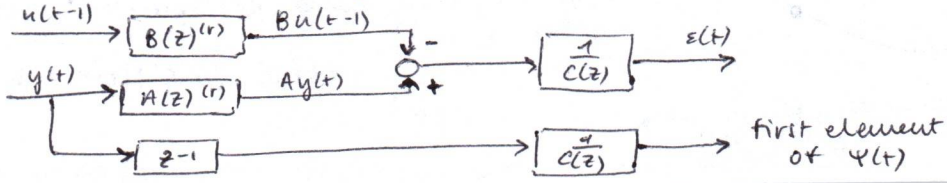
At iteration r we can compute $A(z), B(z), C(z)$ by means of the estimated vector ϑ at iteration r.

$\psi(t)$: (from data)

$$\psi(t) = \Big[\underbrace{-\frac{\partial \varepsilon(t)}{\partial a_1}, \dots,}_{\text{derivatives w.r.t. parameters } a_i} \underbrace{-\frac{\partial \varepsilon(t)}{\partial b_1}, \dots,}_{\text{der. w.r.t. param } b_i} \underbrace{-\frac{\partial \varepsilon(t)}{\partial c_1}, \dots}_{\text{" } c_i} \Big]^T$$

$$(*) \Longrightarrow C(z)\left[-\frac{\partial \varepsilon(t)}{\partial a_1} \right] = y(t-1) \longrightarrow -\frac{\partial \varepsilon(t)}{\partial a_1}$$

$$\begin{cases} A(z)\,y(t) = y(t) - a_1 y(t-1) - a_2 y(t-2) - \dots \\ \dfrac{\partial A(z)\,y(t)}{\partial a_1} = -y(t-1) \end{cases}$$



The same for the other parameters :

$$C(z)\left[-\frac{\partial \varepsilon(t)}{\partial a_2}\right] = y(t-2) \longrightarrow [\psi(t)]_2$$

Summing up: one can find $\varepsilon(\cdot)$ and all the elements of gradient $\psi(\cdot)$ from data by filtering data with filters computed by using the last estimate of the parameters.

From this, one can pass from $\theta^{(r)}$ to $\theta^{(r+1)}$ with the Gauss-Newton formula.

Then, with $\theta^{(r+1)} \longrightarrow$ repeat up to some convergence.

Note: there is no guarantee to achieve the global min (we do this starting from different points, we achieve different minima and then we chose the min among minima (still no guarantee)).

The estimate $\hat{\theta}_N$, for large values of N, is close to $\bar{\theta}$. $\Rightarrow$ we can study the performance of PEM by analyzing the property of model $M(\bar{\theta})$ (ASYMPTOTIC ANALYSIS)

:= Asymptotical analysis of PEM (prediction error methods)

Example (1) : S : $y(t) = a^o y(t-1) + \xi(t)$ AR(1)

data generated system

$\hat{M}(\theta)$: $\hat{y}(t) = a y(t-1)$ AR(1)

family of models in predictive form

Estimate the model's parameter a with LS by solving the normal equations

$\Longrightarrow \hat{a}_N$

(estimated with N data)

Question : Do $\hat{a}_n \longrightarrow a^o$ as $N \to \infty$?

If the convergence holds $\Rightarrow$ the estimate is consistent

$$J_N(\theta) = \frac{1}{N} \sum_{t=1}^{N} \varepsilon(t)^2 \longrightarrow \bar{J}(\theta) = \mathbb{E}\left[\varepsilon_\theta(t)^2\right]$$

$$\Longrightarrow \underset{\theta}{\arg\min}\,(J_N(\theta)) \longrightarrow \underset{\theta}{\arg\min}\,(\bar{J}(\theta))$$

$$(\hat{\theta}_N) \qquad\qquad (\bar{\theta})$$

Computation of $\bar{J}(\theta)$:

$$\varepsilon_\theta(t) = y(t) - \hat{y}_\theta(t) = y(t) - a y(t-1)$$

$$\begin{aligned}
\bar{J}(\theta) &= \mathbb{E}\left[(y(t) - a y(t-1))^2\right] \\
&= \mathbb{E}\left[y(t)^2 + a^2 y(t-1)^2 - 2a\, y(t)\, y(t-1)\right] \\
&= \mathbb{E}\left[y(t)^2\right] + a^2\, \mathbb{E}\left[y(t-1)^2\right] - 2a\, \mathbb{E}\left[y(t)\, y(t-1)\right] \\
&= \gamma_y(0) + a^2 \gamma_y(0) - 2a\, \gamma_y(1) \\
&= (1+a^2)\, \gamma_y(0) - 2a\, \gamma_y(1)
\end{aligned}$$

$$\min \bar{J}(\theta) \Longrightarrow \frac{d\bar{J}(\theta)}{d\theta} = \frac{d\bar{J}(a)}{da} = 2a\, \gamma_y(0) - 2\gamma_y(1) = 0$$

$$\Longrightarrow \bar{a} = \frac{\gamma_y(1)}{\gamma_y(0)}$$

(estimate of parameter of AR(1) obtained with LS for $N \to \infty$)

Is $\bar{a}$ close to the true parameter of the data gen. system?

If S: AR(1) with parameter a^o then we know that :

$$\gamma(1) = a\, \gamma(1) \Longrightarrow \bar{a} = \frac{\gamma(1)}{\gamma(0)} = a^o$$

$$\Longrightarrow \hat{a}_N \xrightarrow{N \to \infty} \bar{a} = a^o$$

$\Longrightarrow$ the estimate is consistent.

Example $^{(2)}$: $S: \quad y(t) = \xi(t) + c\,\xi(t-1)$ $\qquad$ MA(1)

$\hat{M}(\theta): \quad \hat{y}(t) = a\,y(t-1)$ $\qquad$ (same as example $^{(1)}$)

Compute the optimal estimate:

$$\hat{a}_N \longrightarrow \bar{a} = \frac{\gamma(1)}{\gamma(0)} = \frac{c\lambda^2}{(1+c^2)\lambda^2} = \frac{c}{1+c^2}$$

Important comment:

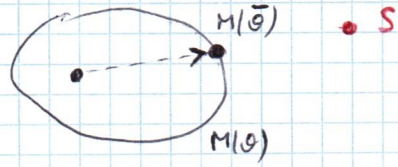
The family of models <u>may</u> encompass the data generation mechanism <u>or not</u>.

$(1)^{st}$ example: $\quad S: \; AR(1)$

$M(\theta): \; AR(1)$

$(2)^{nd}$ example: $\quad S: \; MA(1)$

$M(\theta): \; AR(1)$

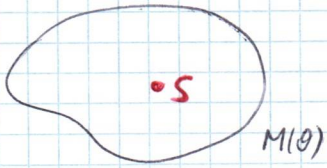


In this case the class of model does not describe perfectly the system generating data <u>(it's only an approximation data)</u>

$M(\bar{\theta})$ is the optimal proxy of S in the sense of square errors.

Asymptotic behaviour of PEM

Suppose that S belongs to the family of models



$$S = M(\theta^\circ)$$

We have studied the problem of convergence of $\hat{\theta}_N$ to θ°.
Can we say something about the rate of convergence?

$$\varepsilon_\theta(t) \Longrightarrow \psi(t) := -\left.\frac{\partial \varepsilon_\theta(t)}{\partial \theta}\right|_{\theta=\theta^\circ} \Longrightarrow \underbrace{\bar{R} := \mathbb{E}\left[\psi(t)\,\psi(t)^T\right]}_{\text{square matrix } n\times n}$$

where n is the number of parameters to be estimated $(n = \dim(\theta))$

$$\Longrightarrow \underbrace{\mathrm{Var}(\hat{\theta}_N - \theta^\circ)}_{\text{uncertainty in the parameters estimation}} \simeq \frac{1}{N}\,\bar{R}^{-1}\,\lambda^2$$

variance of the prediction error evaluated for the true system

$$\lambda^2 = \mathrm{Var}\left(\varepsilon_\theta(t)\big|_{\theta^\circ}\right)$$

In the LS case : $\quad Var(\hat{\theta}_N - \theta^o) = \frac{1}{N} \mathbb{E}[\phi(t)\,\phi(t)^T]\,\lambda^2$

In practice we replace expectation with their sample surrogate :

$$\Longrightarrow \quad Var(\hat{\theta}_N - \theta^o) = \frac{1}{N}\left(\frac{1}{N}\sum_{t=1}^{N} \phi(t)\,\phi(t)^T\right)^{-1}\lambda^2$$

$$\lambda^2 = \frac{1}{N}\sum_{t=1}^{N} \varepsilon_\theta(t)^2$$

LS ESTIMATE PROCEDURE

N data : $u(1), \dots, u(\ )$
$y(1), \dots, y(N)$

$$u \longrightarrow \boxed{ARX(n_a, n_b)} \longrightarrow y$$

$$\Longrightarrow \quad \phi(N) = \begin{bmatrix} y(t-1) \\ \vdots \\ y(t-N) \\ u(t-1) \\ \vdots \\ u(t-N) \end{bmatrix} \quad \Longrightarrow \quad \bar{R}_N = \frac{1}{N}\sum_{t=1}^{N} \phi(t)\,\phi(t)^T$$

square matrix $n \times n$ with $n = n_a + n_b$

Is $\bar{R}_N$ invertible ?

- Yes $\Longrightarrow$ we compute the LS estimate of parameters θ from data :

$$\hat{\theta}_N = \bar{R}_N^{-1}\left(\frac{1}{N}\sum_{t=1}^{N} \phi(t)\,y(t)\right) \qquad \text{LS formula}$$

estimate of the optimal parameter in the family of models

Then we can evaluate the prediction error :

$$\varepsilon(t) = y(t) - \phi(t)^T \hat{\theta}_N$$

Is $\varepsilon(\cdot) \sim WN$? Suppose yes (for checking it we compute $\gamma(\tau)$ and we check that is like :)

$$\Longrightarrow \quad \lambda^2 = \frac{1}{N}\sum_{t=1}^{N}\left(y(t) - \phi(t)^T \hat{\theta}_N\right)^2$$

estimate of the variance of WN affecting the LS model (ARX model)

$$\Longrightarrow \quad Var(\hat{\theta}_N - \theta^o) = \begin{array}{c} \left[\begin{array}{cc|c} \lambda_{11} & & \\ & \lambda_{22} & \\ & & \ddots \\ \hline & & \end{array}\right] \end{array} \begin{array}{l} \}\, n_a \\ \}\, n_b \end{array}$$

where :
- λ_{11} := variance of the error of the first par. of θ (a_1)
- ..
- λ_{kk} := " " θ (a_k)

(then we judge if we're in a good condition or not)

- No $\Longrightarrow$ we can study the possibility of changing the family of models

Example: Identification of an ARX model with different order from data.

$$S: \quad y(t) = 1.2\, y(t-1) - 0.32\, y(t-2) + u(t-1) + 0.52\, u(t-2) + \xi(t)$$

$$\xi(t) \sim WN(0,1)$$

$$u(t) \sim WN(0,4)$$

$$\hat{M}(\theta): \hat{y}(t) = a_1 y(t-1) + \dots + a_{n_a} y(t-n_a) + b_1 u(t-1) + \dots + b_{n_b} u(t-n_b)$$

$(ARX(n_a, n_b))$

Cases:
- $n_a = n_b = 1$
- $n_a = n_b = 2$
- $n_a = n_b = 3$

We generate 2000 snapshots ($\equiv$ samples) of y from 2000 snapshots of u. We use the LS procedure to find optimal ARX model in the cases above.
Results obtained:

- ARX (1,1): $\hat{a}_1 = 0.93$ (0.6%)
  $\hat{b}_1 = 0.97$ (2.3%)
  $J = 3.86$
  Anderson test: NO
- ARX (2,2): $\hat{a}_1 = 1.204$ (1%) $\hat{b}_1 = 0.904$ (1%)
  $\hat{a}_2 = -0.320$ (3%) $\hat{b}_2 = 0.485$ (3%)
  $J = 0.99$
  Anderson test: OK
- ARX (3,3): $\hat{a}_1 = 1.19$ (2%) $\hat{b}_1 = 0.98$ (1%)
  $\hat{a}_2 = -0.299$ (10%) $\hat{b}_2 = 0.494$ (5%)
  $\hat{a}_3 = -0.019$ (68%) $\hat{b}_3 = -0.016$ (120%)
  Anderson test: OK

$\Longrightarrow$ ARX (2,2) is the best

MODEL COMPLEXITY SELECTION

We want to estimate the optimal AR in a family of AR(n) for $n = n_1, \dots, n_2$ among all possible orders from n_1 to n_2.

Which is the best order?

First idea: estimate best model AR(n) with LS $\hat{\theta}_N$ best $M(\hat{\theta}_N)$.
then we can compute

$$\varepsilon_{\hat{\theta}_N}(t) = y(t) - \hat{y}_{\theta_N}(t) \longrightarrow \underbrace{J = \sum_t \varepsilon_{\hat{\theta}_N}(t)^2}_{\text{performance of fit of the model}}$$

We can chose the optimal order of an AR model by trying to minimize such J.

However, it is <u>not</u> a good idea.

So :

We have a data sequence : $y(1), y(2), \dots, y(\ )$.

We would like to fit an AR model to such data by a full black-box approach :

$$\underbrace{\hat{y}(t)}_{\text{model of order } n} = a_1 y(t-1) + \dots + a_n y(t-n)$$

$$\underset{\text{parameter vector}}{\theta} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$$

$$\underset{\text{observation vector}}{\phi(t)} = \begin{bmatrix} y(t-1) \\ \vdots \\ y(t-n) \end{bmatrix}$$

$\Longrightarrow$ model : $\hat{y}(t) = \phi(t)^T \theta$

$\Longrightarrow$ prediction error of the model $M(\theta)$: $\varepsilon_\theta(t) = y(t) - \phi(t)^T \theta$

We can estimate θ by LS :

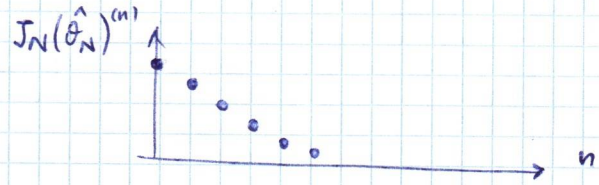
$$\min_\theta J_N(\theta) = \min_\theta \frac{1}{N} \sum_{t=1}^{N} \varepsilon_\theta(t)^2$$

$$= \min_\theta \frac{1}{N} \sum_{t=1}^{N} (y(t) - \phi(t)^T \theta)^T$$

$\Longrightarrow \hat{\theta}_N$

After computing $\hat{\theta}_N$ we can find $\boxed{J_N(\hat{\theta}_N)^{(n)}}$ ($\equiv$ for models of order n)

= index of fit of the model estimated with a given sequence of data on the same data used in identification :



$J_N(\hat{\theta}_N)^{(n)}$ is monotonically decreasing with n (when n increase we have more degrees of freedom in the model and we can obtain a better fit)

How to chose n? How to chose the BEST ORDER ($\equiv$ BEST COMPLEXITY)?

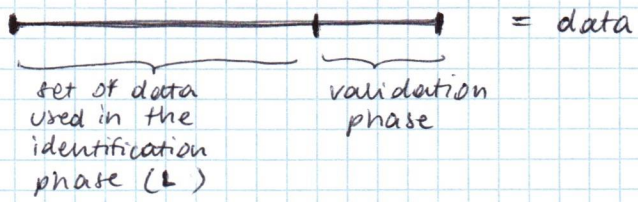
When n is too large there is a serious risk of fitting also the noise. (OVERFITTING).

Main point : if we take $J_N(\hat{\theta}_N)^{(n)}$ for evaluation of fit, then the data are used :

- to find $\hat{\theta}_N$ (LS)
- to compute J_N

$\Longrightarrow$ it's a subjective fitting

How can we set up an objective fitting?
We can partitionate the data:



We use the first set of data for identification and the second set of data for a fair evaluation of J_N (it corresponds to use fresh data).
Then we minimize J_N evaluated over validation data w.r.t. n.

(This idea is called CROSS VALIDATION)

However, there is a main problem: what can we do for sequencies of data not as long? Can we find an objective index of fitting from a subjective one such as $J_N(\hat{\theta}_N)^{(n)}$?

Yes, we can.
We have 3 main criteria:

1. FPE (Final prediction error)
2. AIC (Akaike Information criteria)
3. MDL (Minimum description length)

FPE (Final Prediction Error)

We have a sequence of data : $y(t)$.
(Note that this sequence of data is the outcome of an underlying random event s, so we should write $y(t,s)$).

$$\Longrightarrow \quad \bar{J}(\theta) = \underbrace{\mathbb{E}\left[\left(y(t,s) - \hat{y}(t,s,\theta)\right)^2\right]}_{\text{average fitting overall possible outcomes of the underlying experiments (} \equiv \text{ average overall possible sequencies of data)}} \qquad (\theta \text{ fixed})$$

We want to replace θ with $\hat{\theta}_N(s)$ (which depends on data (s))

$$\theta \longrightarrow \hat{\theta}_N(s) \quad \Longrightarrow \quad \bar{J}(\theta) \longrightarrow \bar{J}(\hat{\theta}_N(s))$$

Def. $\quad FPE = \mathbb{E}_s\left[\bar{J}_N(\hat{\theta}_N(s))\right]$

average accuracy of all possible estimates obtained with all possible sequencies of data

$\Longrightarrow$ FPE is an objective index of fit

Can we compute FPE from data? Yes, we can.

Computing FPE:

Assumptions: assume that data are generated by:

$$S: \quad y(t) = \phi(t,s)^T \theta^0 + \eta(t) \qquad \eta(t) \sim WN(0, \lambda^2)$$

We use the family of models:

$$\hat{M}(\theta): \quad y(t,s,\theta) = \phi(t,s)^T \theta$$

in predictive form

fixed

model provided by a given vector θ of parameter

$$\phi(t,s) = \begin{bmatrix} y(t-1,s) \\ \vdots \\ y(t-n,s) \end{bmatrix}$$ = sequence of data associated with a given s

(NOTE: it's a function of time up to $t-1$ and go on up to $t-n$)

$$\Longrightarrow \mathbb{E}[\phi(t,s)\,\eta(t)] = \mathbb{E}[\phi(t,s)]\,\mathbb{E}[\eta(t)] = 0$$

$\phi(t,s)$ depends on the past $\Rightarrow$ $\phi(t,s) \perp \eta(t)$

I step: compute $\bar{J}(\theta)$:

$$\begin{aligned}
\bar{J}(\theta) &= \mathbb{E}[(y(t,s) - \hat{y}(t,s,\theta))^2] \\
&= \mathbb{E}[(\phi^T\theta^0 + \eta - \phi^T\theta)^2] \\
&= \mathbb{E}[(\phi^T(\theta^0 - \theta) + \eta)^2] \\
&= (\theta^0-\theta)^T\,\mathbb{E}[\phi(t,s)\,\phi(t,s)^T]\,(\theta^0-\theta) + \lambda^2 + \underbrace{\cancel{\mathbb{E}[\ \cdot\ ]}}_{\phi(t,s)\,\perp\,\eta(t)} \\
&= (\theta^0-\theta)^T\,\bar{R}\,(\theta^0-\theta) + \lambda^2 \qquad (*)
\end{aligned}$$

(where $\bar{R} = \mathbb{E}[\phi\phi^T]$, square $n \times n$ matrix)

II step: Recall: in LS $\quad \mathrm{Var}((\theta^0 - \hat{\theta}_N(s)) \simeq \frac{\lambda^2}{N}\bar{R}^{-1}$

$(\theta^0 - \hat{\theta}_N(s)) := \nu$

We now replace in $(*)$ θ with $\hat{\theta}_N(s)$.

$$\Longrightarrow \mathrm{FPE} = \mathbb{E}[\nu^T\,\mathrm{Var}(\nu)^{-1}\,\nu]\,\frac{\lambda^2}{N} + \lambda^2$$

Nice formula: $\mathbb{E}[\nu^T\,\mathrm{Var}(\nu)^{-1}\,\nu] = \dim(\nu) = n$

$\boxed{}\ \boxed{}\ \boxed{}$ = scalar

$$\Longrightarrow \mathrm{FPE} = \frac{n}{N}\lambda^2 + \lambda^2 = \frac{N+n}{N}\lambda^2$$

variance of η

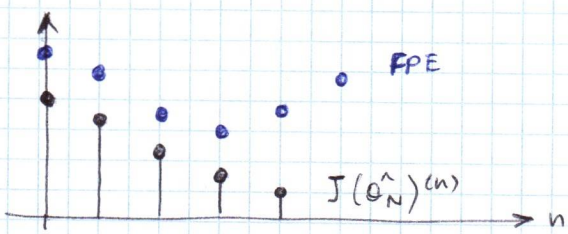
III step: We have to estimate λ^2 from data:

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

$$\Longrightarrow \mathrm{FPE} = \left(\frac{n}{N} + 1\right)\frac{N}{N-n}\,J(\hat{\theta}_N)^{(n)}$$

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

$\Rightarrow$ the optimal complexity is that for which FPE is minimal



(Recap of the) Procedure to estimate optimal order :

- we have data
- we consider the AR model of order n
- with LS estimate we estimate the par. vector θ : $\hat{\theta}_N$
- we compute the subjective index of fitting :

$$J(\hat{\theta}_N)^{(n)} = \frac{1}{N} \sum \varepsilon_{\hat{\theta}_N}(t)^2$$

- repeat for a set of values of n $(n = 1, 2, \dots, 30)$
- compute :

$$FPE = \frac{N+n}{N-n} J$$

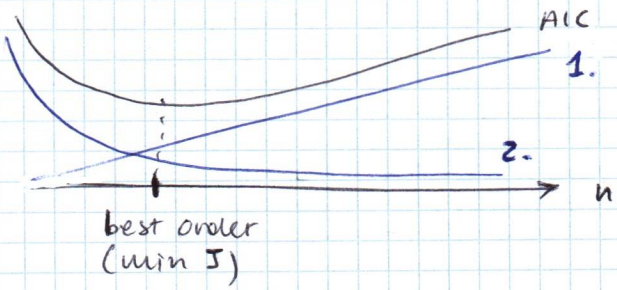
- find an optimal order n by min FPE

AIC (Akaike Information Criteria)

Def. $$AIC = \underbrace{\ln(N)\frac{n}{N}}_{1.} + \underbrace{\ln(J)}_{2.}$$

1. penalization of complexity (increasing with n)
2. decreasing with order n

- fix n
- calculate $\hat{\theta}_N$
- evaluate $J(\hat{\theta}_N)$
- repeat for some n
- create the AIC curve



Comparison : AIC vs. FPE

$$FPE = \frac{N+n}{N-n} J = \frac{1+\frac{n}{N}}{1-\frac{n}{N}} J$$

$$\Longrightarrow \ln(FPE) = \ln\left(1+\frac{n}{N}\right) - \ln\left(1-\frac{n}{N}\right) + \ln(J)$$

We know that $\ln(1+x) \sim x$ when x is very small,

since $n \ll N$: $\ln\left(1+\frac{n}{N}\right) \sim \frac{n}{N}$

$\ln\left(1-\frac{n}{N}\right) \sim -\frac{n}{N}$

$$\Longrightarrow \ln(FPE) = 2\frac{n}{N} + \ln(J) \equiv AIC$$

Conclusion : for large number of data N, AIC is equivalent to FPE (same estimate of the optimal order)

MDL (Minimum description length)

= binary coding approach

We have to code data $y(1), y(2), \dots, y(N)$ in binary code (or alternatively we can code the model and prediction error ($M(\theta)$ and $\varepsilon(1), \varepsilon(2), \dots, \varepsilon(N)$))

$$y(1), \dots, y(N) \equiv M(\theta) + \varepsilon(1), \dots, \varepsilon(N)$$

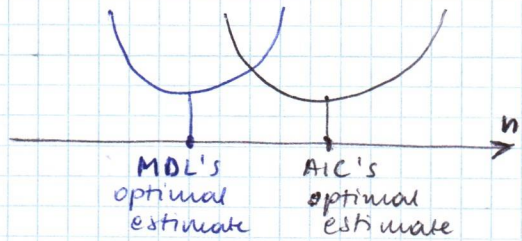
What is the effort? The binary coding of θ and $\varepsilon(1), \dots, \varepsilon(N)$.
If we increase the model order (n), θ becomes longer and so, the effort of binary coding increase with n. On the opposite, by increasing n, the prediction error becomes smaller.

$$\boxed{MDL = \ln(N)\frac{n}{N} + \ln(J)}$$

$\longrightarrow$ we want to find a compromise between the effort in coding and prediction error

$$AIC = 2\ln(N)\frac{n}{N} + \ln(J)$$

The penalization of complexity n is higher in MDL (w.r.t. AIC)
$\Longrightarrow$ minimizing MDL w.r.t. complexity leads to lower complexities:



$\Longrightarrow$ MDL leads to more parsimonious models than AIC (less number of parameters)

DURBIN-LEVINSON ALGORITHM

Goal: estimate AR(n) models of various order : $n = 1, 2, \dots$. At each step we use the LS formula ($\Leftrightarrow$ matrix inversion at each step).
Is it possible to estimate the parameters of an AR(n+1) from the estimate of model AR(n)?

Is there a recursion to pass from AR(n) to AR(n+1)?

Yes, the Durbin Levinson formula

Focus on: AR(1) $\longrightarrow$ AR(2)

AR(1): $y(t) = \underline{a_1}\, y(t-1) + \underline{\eta(t)}$ $\qquad \eta(t) \sim WN(0, \underline{\lambda^2})$

AR(2): $y(t) = \underline{a_1}\, y(t-1) + \underline{a_2}\, y(t-2) + \underline{\eta(t)}$ $\qquad \eta(t) \sim WN(0, \underline{\lambda^2})$

(Note that the — set is different from the — set).

We can estimate the parameters performing an LS or solving the Yule-Walker equations (it's the same for a data large enough).

$$\Longrightarrow \begin{cases} a_2 = \frac{1}{\lambda^2}\left(\gamma(2) - a_1\,\gamma(1)\right) \\ a_1 = a_1 - a_2\frac{\gamma(1)}{\gamma(0)} \\ \lambda^2 = \lambda^2\left(1 - a_2^2\right) \end{cases}$$

DURBIN LEVINSON for AR(1) $\longrightarrow$ AR(2)

Remark: if the data generation mechanism is AR(1)

$\Longrightarrow \gamma(2) = a_1 \gamma(1)$

$\Longrightarrow a_2 = 0, \quad a_1 = a_1, \quad \lambda^2 = \lambda^2$

$\Longrightarrow$ AR(2) downgrade to AR(1)
(Durbin-Levinson tells us to go back to AR(1))

Remark: the D-L recursion can be generalized to: $AR(n) \rightarrow AR(n+1)$

TIME SERIES ANALYSIS MADE EASY

Given a time series (sequence of real data): $y(1), y(2), \dots, y(N)$, we want to fit a suitable model.

Possible models:

- MA(k) $\quad k = 1, 2, \dots, n$
- AR(k) $\quad k = 1, 2, \dots, n$
- ARMA(k_a, k_b) $\quad k_a = 1, 2, \dots, n_a$, $k_b = 1, 2, \dots, n_b$

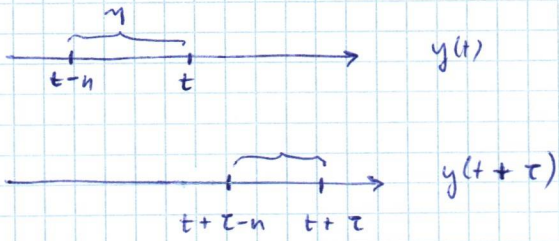
We want to find the best family and the most appropriate order (k, k_a, k_b).

We can estimate ARMA models (with the so-called Maximum Likelihood (ML) method) for various orders and then apply the model selection criteria: FPE, AIC, MDL for best choice.

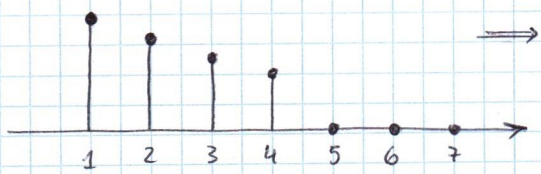
- **MA(n)**: $y(t) = \eta(t) + c_1 \eta(t-1) + \dots + c_n \eta(t-n) \qquad \eta \sim WN(0, \lambda^2)$

Assuming that is a stationary process:

$$\gamma(\tau) = 0 \qquad |\tau| \geq n+1$$



From data we can compute $\gamma(\tau)$ or its approximated expression $\hat{\gamma}(\tau)$:



$\Longrightarrow$ in this case the suitable order is 4

The covariance function $\gamma(\tau)$ can be very usefull to detect the appropriate order of a MA model to fit data.

- **AR(n_a)**: $y(t) = a_1 y(t-1) + \dots + a_{n_a} y(t-n_a) + \eta(t) \qquad \eta \sim WN(0, \lambda^2)$

Model parameters:

$$\underset{\text{param. vector}}{\theta} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{n_a} \end{bmatrix} \qquad \underset{\text{observation vector}}{\phi} = \begin{bmatrix} y(t-1) \\ \vdots \\ y(t-n_a) \end{bmatrix}$$

$$\Longrightarrow \hat{y}(t) = \phi(t)^T \theta$$

(AR model in predictive form)

The model is linear in ϑ.

We estimate $\hat{\vartheta}$ with **LS** or **YW**. It requires inverting a $n_a \times n_a$ matrix.

Moving from n_a to n_a+1 requires inverting a $(n_a+1)\times(n_a+1)$ matrix, so to estimate AR models of order – say – 1, 2, .., 100, it's necessary a lot of inversion of square matrices of various dimensions.
This can be avoided with DURBIN-LEVINSON RECURSION (easy update of the parameters of an AR model of order n to find the parameters of an AR model of order $n+1$)

AR(1): $y(t) = a_1 y(t-1) + \eta(t)$ $\qquad \eta \sim WN(0, \lambda^2)$

AR(2): $y(t) = a_1 y(t-1) + a_2 y(t-2) + \eta(t)$ $\qquad \eta \sim WN(0, \lambda^2)$

AR(3): $y(t) = a_1 y(t-1) + a_2 y(t-2) + a_3 y(t-3) + \eta(t)$ $\qquad \eta \sim WN(0, \lambda^2)$

$\vdots$

Problem of estimating parameters from data: all the a_1 (for the models AR(1), AR(2), ...) are different, $\forall a_i$.

$\Rightarrow$ AR(k): $y(t) = a_1^{(k)} y(t-1) + \dots + a_k^{(k)} y(t-k) + \eta(t)$ $\qquad \eta \sim WN(0, \lambda^{2(k)})$

let's focus on the diagonal elements $(a_1^{(1)}, a_2^{(2)}, \dots, a_k^{(k)})$:

AR(1): $y(t) = a_1^{(1)} y(t-1) + \eta(t)$ $\qquad \eta \sim WN(0, \lambda^{2(1)})$

AR(2): $y(t) = a_1^{(2)} y(t-1) + a_2^{(2)} y(t-2) + \eta(t)$ $\qquad \eta \sim WN(0, \lambda^{2(2)})$

AR(3): $y(t) = a_1^{(3)} y(t-1) + a_2^{(3)} y(t-2) + a_3^{(3)} y(t-3) + \eta(t)$ $\qquad \eta \sim WN(0, \lambda^{2(3)})$

$\vdots$

The parameters $a_k^{(k)}$ have special meaning.

Note that $\forall k \; |a_k^{(k)}| < 1$

For example: AR(2)

$$y(t) = a_1 z^{-1} y(t) + a_2 z^{-2} y(t) + \eta(t)$$

$$(1 - a_1 z^{-1} - a_2 z^{-2})\, y(t) = \eta(t)$$

$$\Rightarrow W(z) = \frac{1}{1 - a_1 z^{-1} - a_2 z^{-2}} = \frac{z^2}{z^2 - a_1 z - a_2}$$

Since the process is stationary, $W(z)$ needs to be stable ($\Rightarrow$ the singularities of $W(z)$ must be inside the unit disc)

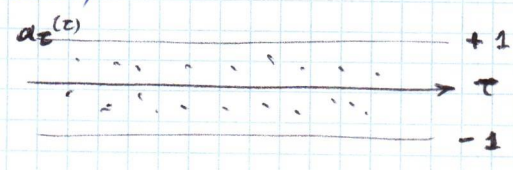
$\Rightarrow$ a_2 = product of all these singularities

$\Rightarrow$ $|a_2| < 1$

Iteratively, the last parameter of an AR(k) $(a_k^{(k)})$ must satisfy the condition $|a_k^{(k)}| < 1$.

Def. Partial covariance function: $PARCOV(\tau) := a_\tau^{(\tau)}$
(last coefficient of an AR model)

Note that $|a_\tau^{(\tau)}| < 1$:

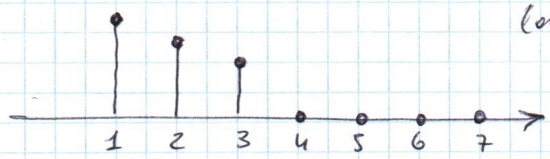


If data are generated by an AR ($n°$) system, then:

$PARCOV(\tau) = 0 \quad \forall \tau > n°$

$\Longrightarrow$ We estimate the $PARCOV(\tau)$ from data for various τ:

(diagram of $PARCOV(\tau)$)



$\Longrightarrow$ We can conclude that data are generated by AR(3). By analizing the $PARCOV(\tau)$ we can see if the data were generated by an AR(k) with k fixed.

So, **RECAP** (TIME SERIES ANALYSIS)

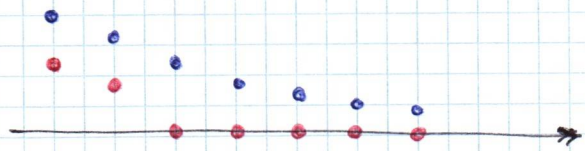
From data we construct: $\longrightarrow$ covariance function $\gamma(\tau)$
$\longrightarrow$ partial covariance function $PARCOV(\tau)$

Remark: we can def. normalized cov. function $\rho(\tau)$:

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)}$$

Since $|\gamma(\tau)| < \gamma(0) \Longrightarrow |\rho(\tau)| < 1$

$\Longrightarrow$ Both the normalized cov. function and the partial covariance function are in the $-1, +1$ strip.



- normalized cov. function
- PARCOV (easy computable with DURBIN-LEVINSON algorithm).

$\Longrightarrow$ AR (2)

RECURSIVE IDENTIFICATION METHODS

15/04

We always considered BATCH METHODS (all data are used together)

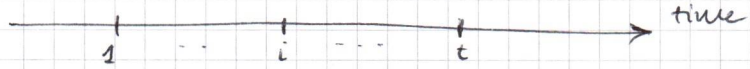
LS $\longrightarrow$ RLS
RECURSIVE LEAST SQUARE

Basic least square formula
(solution of the normal equation)

$$\hat{\theta}_t := S(t)^{-1} \sum_{i=1}^{t} \phi(i)\, y(i)$$

(LS formula at time t)
LS at time t−1 :

$$\hat{\theta}_{t-1} = S(t-1)^{-1} \sum_{i=1}^{t-1} \phi(i)\, y(i)$$



$\phi(i)$:= observation vector
$y(i)$:= observation sequence
$S(t) := \sum_{i=1}^{t} \phi(i)\, \phi(i)^T$:= auxiliary matrix (square matrix)

Recursive version :

Note that :

$$\underbrace{\sum_{i=1}^{t} \phi(i)\, y(i)}_{1 \div t} = \underbrace{\sum_{i=1}^{t-1} \phi(i)\, y(i)}_{1 \div (t-1)} + \underbrace{\phi(t)\, y(t)}_{t}$$

$$= S(t-1)\, \hat{\theta}_{t-1} + \phi(t)\, y(t)$$

$$\sum_{i=1}^{t} \phi(i)\, y(i) = S(t-1)\, \hat{\theta}_{t-1} + \phi(t)\, y(t) \qquad (*)$$

Auxiliary matrix $S(t)$:

$$S(t) = \sum_{i=1}^{t} \phi(i)\, \phi(i)^T = \underbrace{\sum_{i=1}^{t-1} \phi(i)\, \phi(i)^T}_{S(t-1)} + \phi(t)\, \phi(t)^T$$

$$\Rightarrow \quad S(t-1) = S(t) - \phi(t)\, \phi(t)^T$$

$$\Rightarrow \quad (*) : \quad \sum_{i=1}^{t} \phi(i)\, y(i) = \left(S(t) - \phi(t)\, \phi(t)^T\right) \hat{\theta}_{t-1} + \phi(t)\, y(t)$$

$\rightarrow$ $\hat{\theta}_t = S(t)^{-1} \sum_{i=1}^{t} \phi(i)\, y(i)$ becomes :

$$\hat{\theta}_t = S(t)^{-1} \left(\left(S(t) - \phi(t)\, \phi(t)^T \right) \hat{\theta}_{t-1} + \phi(t)\, y(t) \right)$$

$$= \hat{\theta}_{t-1} + \phi(t) \underbrace{\left[y(t) - \phi(t)^T \hat{\theta}_{t-1} \right]}_{:= \varepsilon(t)} + S(t)^{-1} \hat{\theta}_{t-1}$$

In conclusion, updating estim. eq.

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \phi(t)\, \varepsilon(t) + S(t)^{-1}$$

$$\varepsilon(t) = y(t) - \phi(t)^T$$

Or, analogously:

$$\boxed{\begin{aligned} \hat{\theta}_t &= \hat{\theta}_{t-1} + k(t)\,\varepsilon(t) \\ \varepsilon(t) &= y(t) - \phi(t)^T \hat{\theta}_{t-1} \\ k(t) &= S(t)^{-1}\,\phi(t) \\ S(t) &= S(t-1) + \phi(t)\,\phi(t)^T \end{aligned}}$$

$k(t)$:= GAIN of the algorithm (it decides how to change)

Recursion for $\hat{\theta}$

RLS

Recursion for S

$\varepsilon(t)$: prediction error of model estimated at time $t-1$: $\hat{\theta}_{t-1}$

If $\varepsilon(t) = 0$ the model previously estimated perfectly fits the new snapshots $y(t)$. Consequently:

$$k(t) = 0$$

In other words: $\hat{\theta}_t = \hat{\theta}_{t-1}$

INITIALIZATION ?

$$\underbrace{1 \qquad\qquad t_0}_{\text{data}} \longrightarrow \quad \hat{\theta}_{t_0} = S(t_0)^{-1} \sum_{i=1}^{t_0} \phi(i)\, y(i)$$

$$S(t_0) = \sum_{i=1}^{t_0} \phi(i)\,\phi(i)^T$$

At each step the inverse of an $n \times n$ matrix must be computed, where n is the number of parameters to be estimated.

Question: can we avoid such inversion? Yes, we can.

NICE MATRIX FORMULA:
(Matrix inversion lemma)

Suppose we have 4 matrices: F, G, H, K and we want: $(F + GHK)^{-1}$

$$\Longrightarrow \quad (F + GHK)^{-1} = F^{-1} - F^{-1}G\left(H^{-1} + KF^{-1}G\right)^{-1} KF^{-1}$$

We apply this formula to invert:

$$S(t) = S(t-1) + \phi(t)\,\phi(t)^T \longrightarrow \begin{aligned} F &= S(t-1) \\ G &= \phi(t) \\ H &= 1 \\ K &= \phi(t)^T \end{aligned}$$

$$\longrightarrow \quad S(t)^{-1} = S(t-1)^{-1} - S(t-1)^{-1}\phi(t)\underbrace{\left(1 + \phi(t)^T S(t-1)^{-1}\phi(t)\right)}_{\text{scalar}}{}^{-1}\phi(t)^T S(t-1)^{-1}$$

Then we can define a new auxiliary matrix $v(t) = S(t)^{-1}$ and write RLS with $v(t)$ as auxiliary matrix.

Observation (about ausiliary matrix) :

$$S(t) = \sum_{i=1}^{t} \phi(i)\,\phi(i)^T$$

$$S(t) = S(t-1) + \phi(t)\,\phi(t)^T$$

$\Longrightarrow \quad S(t) \geq S(t-1)$

$\Longrightarrow$ $S(t)$ is an increasing sequence of matrixes

$$V(t) = S(t)^{-1} \Longrightarrow V(t) \leq V(t-1)$$

$\Longrightarrow$ $V(t)$ is a decreasing sequence of matrixes

We expect that gain $k(t) \longrightarrow 0$ hence $\hat{\theta}_t \longrightarrow$ constant.

BASIC POSITION OF LS PROBLEM :

$$J = \frac{1}{N} \sum_{t=1}^{N} (y(t) - \phi(t)\,\theta)^2 \longrightarrow J = \frac{1}{N} \sum_{t=1}^{N} \mu^{N-t} (y(t) - \phi(t)\,\theta)^2$$

$\uparrow$ constant vector

weighting coefficient

We can also transform ML into RML

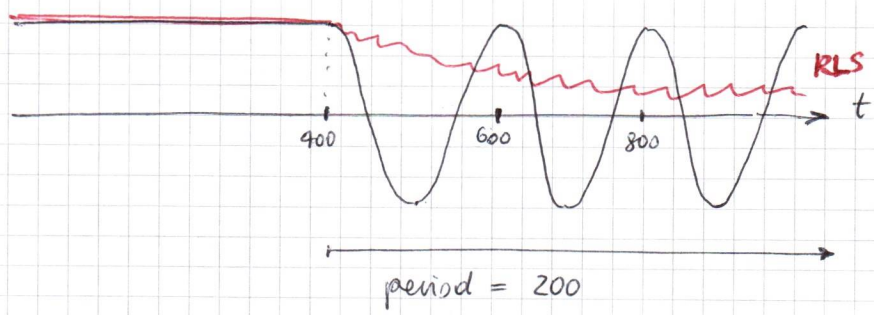
VARYING PARAMETERS

TRACK PARAMETERS

Ex. $y(t) = \theta_1(t)\,u_1(t) + \theta_2(t)\,u_2(t) + \xi(t)$

$\theta_2(t) =$ constant

$\theta_1(t) =$ varying :

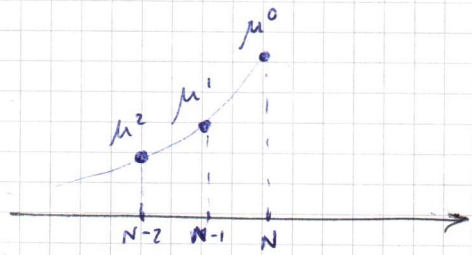


The model is linear in the parameters :

vector of parameters : $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$

μ^{N-t} weighting coefficient :

- $t = N$: $\mu^{N-t} = 1$
- $t = N-1$: $\mu^{N-(N-1)} = \mu$
- $t = N-2$: $\mu^{N-(N-2)} = \mu^2$



the weight is decreasing as t becomes lower and lower

The introduction of the weighting term μ^{N-t} in the least square performance index is most useful to provide more importance to recent data (fresh data) and to forget past data

$\Longrightarrow$ in this way we can track parameter variations

Precisely, with weighting coefficients, RLS modifies as follows:

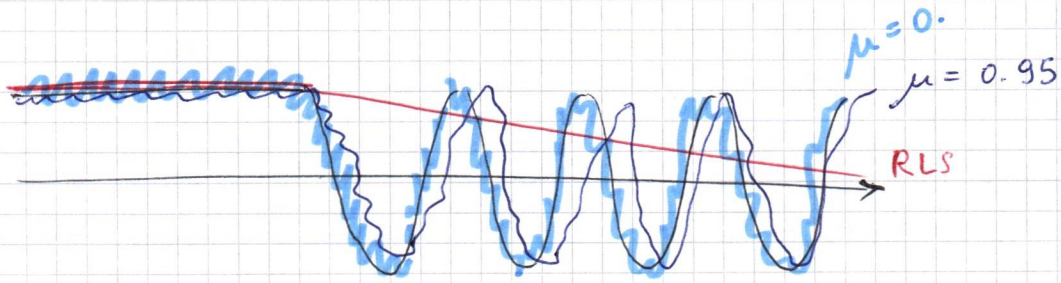
$$\hat{\theta}_t = \hat{\theta}_{t-1} + k(t)\,\varepsilon(t)$$

$$k(t) = S(t)^{-1}\,\phi(t)$$

$$\varepsilon(t) = y(t) - \phi(t)^T \hat{\theta}_{t-1}$$

$$S(t) = \mu\, S(t-1) + \phi(t)\,\phi(t)^T$$

information from data from 1 to t-1. Multiplying by μ we give lower importance to past data w.r.t. the fresh snapshots

:= FORGETTING FACTOR ($\mu < 1$, $\mu > 0$)

Ex.