System Identification
Lecture 4

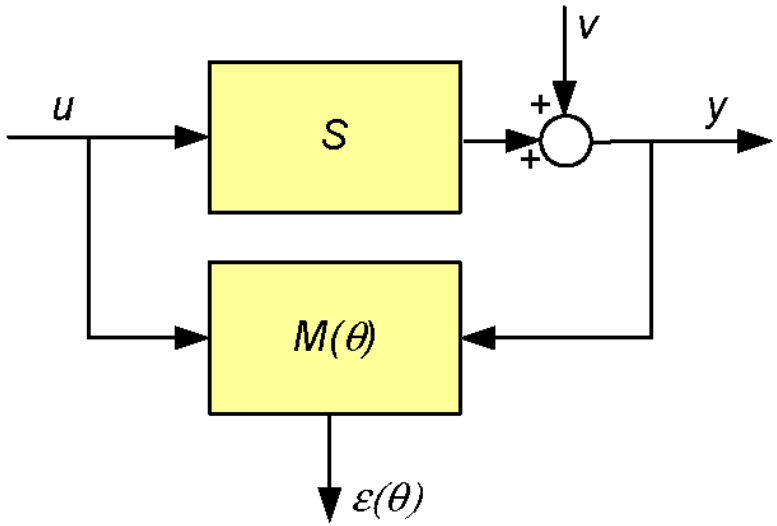
Prediction error method

Paul Van den Hof

Control Systems Group
Department of Electrical Engineering
Eindhoven University of Technology

Contents

Objective:



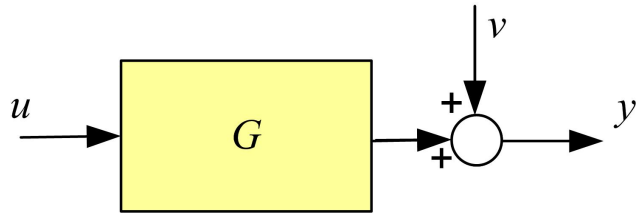
Find an appropriate way to generate a residual signal $\varepsilon(\theta)$ that can serve as a basis for identification with an appropriate identification criterion.

Approach:

- Develop a unique representation of the data generating system (including noise)
- Develop the concept of one-step-ahead prediction
- Define model sets of parametrized predictor models
- Define a justifiable identification criterion

System setup

$$y(t) = G(q)u(t) + v(t)$$



Additive disturbance on the output.

Disturbance signal v:

Measurement noise; non-measurable input signals; effects of linearization; $\cdots$

Noise model:

v is a stationary stochastic process, with $\mathbb{E}v(t) = 0 \quad \forall t$, and a spectral density $\Phi_v(\omega)$ that is smooth (finite dimensional).

Spectral factorization theorem

If the stochastic process v is generated through a finite dimensional system, i.e. $v(t) = H(q)e(t)$ with H a rational transfer function and e a stationary white noise, then there exists a unique decomposition

$$\Phi_v(\omega) = H(e^{i\omega})\sigma_e^2 H(e^{i\omega})^* = \sigma_e^2 \cdot |H(e^{i\omega})|^2$$

with:

- $H(q)$ and $H^{-1}(q)$ are stable,
- H is monic, i.e. $\lim_{z\to\infty} H(z) = 1$,
- $\mathbb{E}e^2(t) = \sigma_e^2$.

We can then write: $v(t) = H(q)e(t)$, with (H, σ_e^2) as a unique representation of the disturbance process v.

$$H(q) = 1 + \sum_{k=1}^{\infty} h(k) q^{-k}$$

$\{e(t)\}$ is sequence of independent, identically distributed (pdf), random variables $\rightarrow$ white noise.

Specific:

$$\begin{aligned} \mathbb{E}e(t) &= 0 \\ \mathbb{E}e(t)e(t-\tau) &= \sigma_e^2 \cdot \delta(\tau) \end{aligned}$$

(does not say anything about probability density function -pdf)

$\{v(t)\}$ is a realization of a stochastic process with properties:

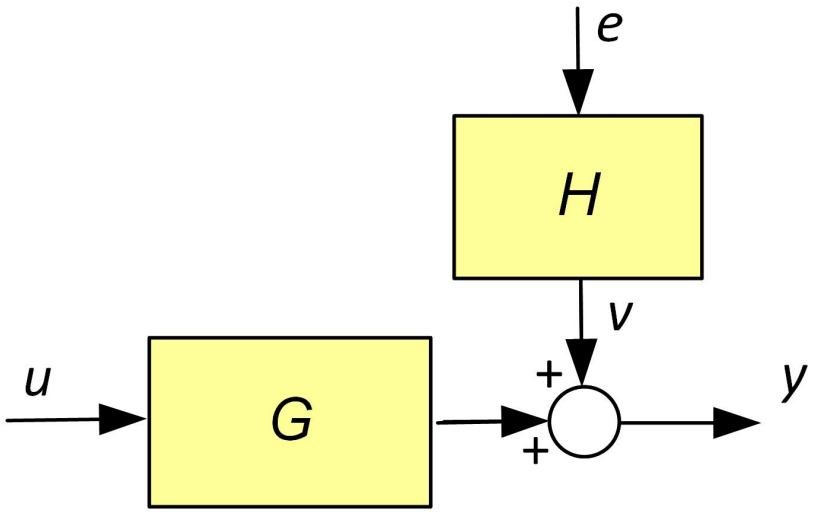
$$\begin{aligned} \mathbb{E}v(t) &= 0 \\ \Phi_v(\omega) &= |H(e^{i\omega})|^2 \cdot \sigma_e^2 \end{aligned}$$

An LTI system/model with (stationary) output noise, is uniquely characterized by the relation

$$y(t) = G(q)u(t) + H(q)e(t)$$

with H monic, stable and stably invertible (minimum-phase).

All poles and zeros of H are within the unit circle.



Prediction

Given a dynamical system

$$y(t) = G(q)u(t) + H(q)e(t)$$

and given observations

$$\{y(s), s \leq t-1; u(s), s \leq t\} \quad i.e. \quad y^{t-1}, u^t$$

how can the future value $\boldsymbol{y(t)}$ be predicted?

Write $y(t)$ as much as possible in terms of past values of y, u:

$$\begin{aligned} y(t) &= G(q)u(t) + H(q)e(t) \\ &= G(q)u(t) + [H(q) - 1]e(t) + e(t) \end{aligned}$$

Substitute in the second term:

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

Then:

$$\begin{aligned} y(t) &= G(q)u(t) + [H(q)-1]H^{-1}(q)[y(t)-G(q)u(t)] + e(t) \\ &= H^{-1}(q)G(q)u(t) + [1 - H^{-1}(q)]y(t) + e(t). \end{aligned}$$

Analyzing the properties of this equation is supported by structural properties of $H^{-1}(q)$.

Property of H^{-1}:

For

$$H(z) = 1 + \sum_{k=1}^{\infty} h(k) z^{-k}$$

there holds (since H is stable and inversely stable)

$$\begin{aligned} \frac{1}{H(z)} &= \frac{1}{1 + h(1)z^{-1} + h(2)z^{-2} + \cdots} \\ &= 1 - h(1)z^{-1} + \cdots \end{aligned}$$

$(1 + \textit{strictly proper})^{-1} = (1 + \textit{strictly proper})$

As a consequence:

- $H^{-1}(q)$ is proper (causal), and
- $1 - H^{-1}(q)$ is strictly proper.

As a result:

$$y(t) = \underbrace{H^{-1}(q)G(q)u(t)}_{\text{known at } t} + \underbrace{[1 - H^{-1}(q)]y(t)}_{\text{known at } t-1} + \underbrace{e(t)}_{\textit{unknown}} .$$

The random variable $y(t)$ conditioned on y^{t-1}, u^t has a pdf that is induced by f_e, with $\mathbb{E}_{f_e} = 0$.
One option for the choice of the "best" predictor is to choose

$$\hat{y}(t|t-1) := \mathbb{E}\{y(t) \mid y^{t-1}, u^t\}$$

which can be calculated on the basis of $\mathbb{E}[e(t)|y^{t-1}, u^t] = 0$.

One-step-ahead prediction

For a dynamical model $y(t) = G(q)u(t) + H(q)e(t)$, with H monic, stable and minimum-phase, the one-step-ahead prediction $\hat{y}(t|t-1) := \mathbb{E}\{y(t)|y^{t-1}, u^t\}$ is given by

$$\hat{y}(t|t-1) = H^{-1}(q)G(q)u(t) + [1 - H^{-1}(q)]y(t)$$

At the same time:

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

The one-step ahead predictor of a dynamic model predicts the output of the model up to the white noise term $e(t)$, called the innovation process.

Prediction in a state space setting (tentative - no exam material)

General state space model:

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) + w(k), \quad x(0) = x_0,\ cov(x_0) = P \\ y(t) &= Cx(t) + Du(t) + v(k) \end{aligned}$$

with w, v white noise processes, can be converted to the innovation form

$$\begin{aligned} \xi(t+1) &= A\xi(t) + Bu(t) + K(t)e(t), \quad \xi(0) = \xi_0 \\ y(t) &= C\xi(t) + Du(t) + e(t) \end{aligned}$$

with

$$\bar{\mathbb{E}}\{y(t) \mid \xi^t, u(t)\} = C\xi(t) + Du(t),$$

and e the innovation process.

The optimal -time-varying- Kalman gain $K(t)$ is obtained through Kalman filter theory.

Solution for $t \to \infty$ is equivalent to transfer function expression.

Example 1 - Predictor for Moving Average (MA) model

$$y(t) = e(t) + ce(t-1)$$

$$H(z) = 1 + cz^{-1}; \qquad H^{-1}(z) = \frac{1}{1 + cz^{-1}}$$

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

So

$$\hat{y}(t|t-1) = \frac{cq^{-1}}{1 + cq^{-1}} y(t)$$

or

$$\begin{aligned} (1 + cq^{-1})\hat{y}(t|t-1) &= cq^{-1}y(t) \\ \hat{y}(t|t-1) + c\hat{y}(t-1|t-2) &= cy(t-1). \end{aligned}$$

Note: $\mathbb{E}y(t) = 0$ but still the prediction of $y(t)$ is unequal 0! (y^{t-1} carries information on $y(t)$).

Example 2 - Predictor for Output error (OE) model

$$y(t) = \frac{b_1 q^{-1}}{1 + a_1 q^{-1}} u(t) + e(t)$$

$$G(q) = \frac{b_1 q^{-1}}{1 + a_1 q^{-1}}; \qquad H(q) = 1.$$

So:

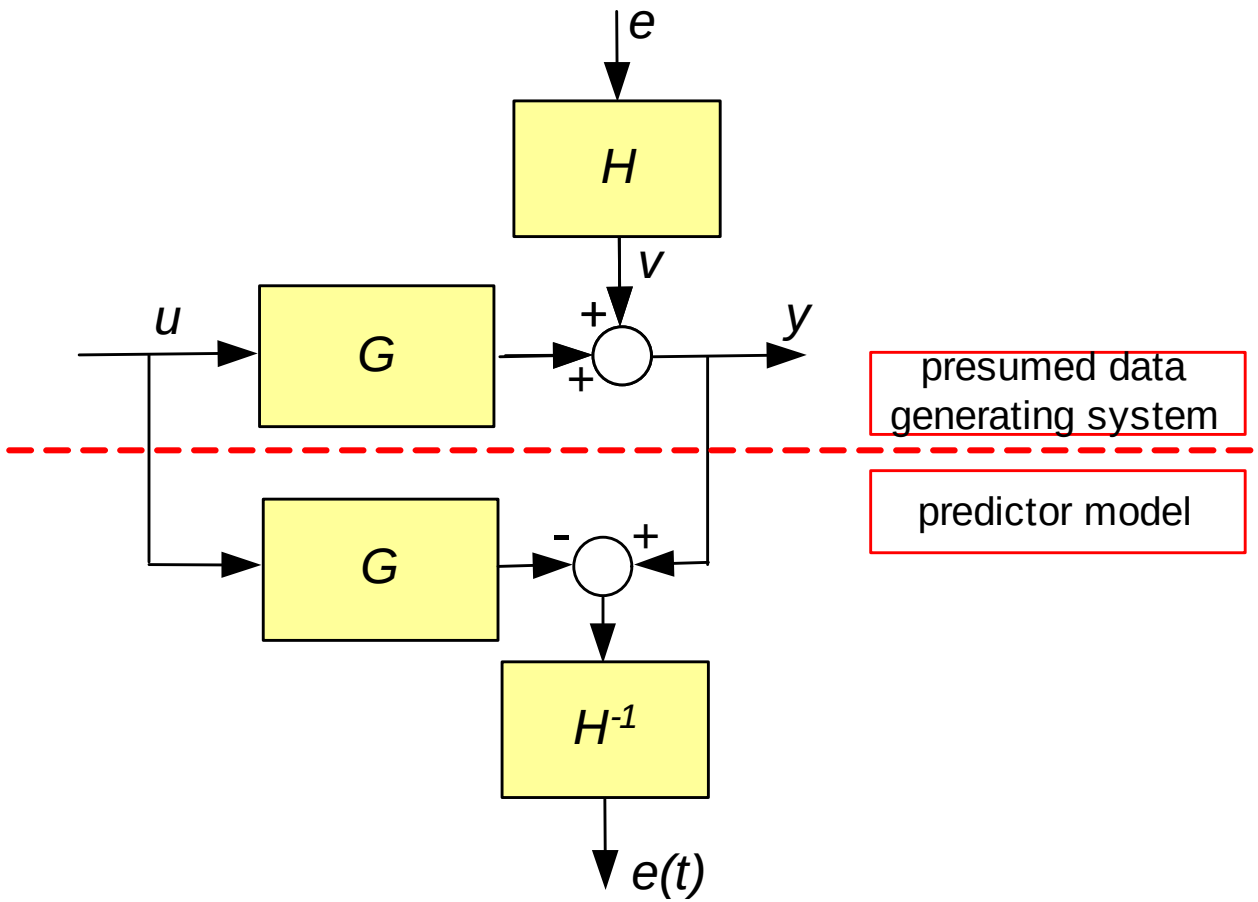
$$\begin{aligned} \hat{y}(t|t-1) &= H^{-1}(q)G(q)u(t) + (1 - H^{-1}(q))y(t) \\ &= G(q)u(t). \end{aligned}$$

Predictor is only driven by input signals.

Note that with

$$\hat{y}(t|t-1) = H^{-1}(q)G(q)u(t) + [1 - H^{-1}(q)]y(t)$$

it follows that $e(t) = y(t) - \hat{y}(t|t-1) = H^{-1}(q)[y(t) - G(q)u(t)]$



Summary

- Framework for linear time-invariant finite-dimensional models with disturbances
- Models are represented by G, H and σ_e^2
- Models induce prediction properties

Prediction errors

The problem in an identification context:

(Unknown) system that has generated the data:

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

with G_0, H_0, e unknown.

Hypothesized model:

$$y_m(t) = G(q)u(t) + H(q)e(t)$$

with e unknown.

Goal of identification:

Find the best estimates G and H on the basis of u and y.

Limitation:
A model can not simulate output data, because of unknown e.

Approach:
A model can predict future outputs on the basis of past input and output.

This motivates the use of predictions as a basis for identification.

One-step-ahead prediction is going to be used as a way to measure the deviation/distance between a data set $\{(y(t), u(t))\}_{t=1,\cdots N}$ and a predictor model $(G(q), H(q))$.

If the data has been generated by (G_0, H_0), then the predictor based on (G_0, H_0) leads to

$$y(t) - \hat{y}(t|t-1) = e(t).$$

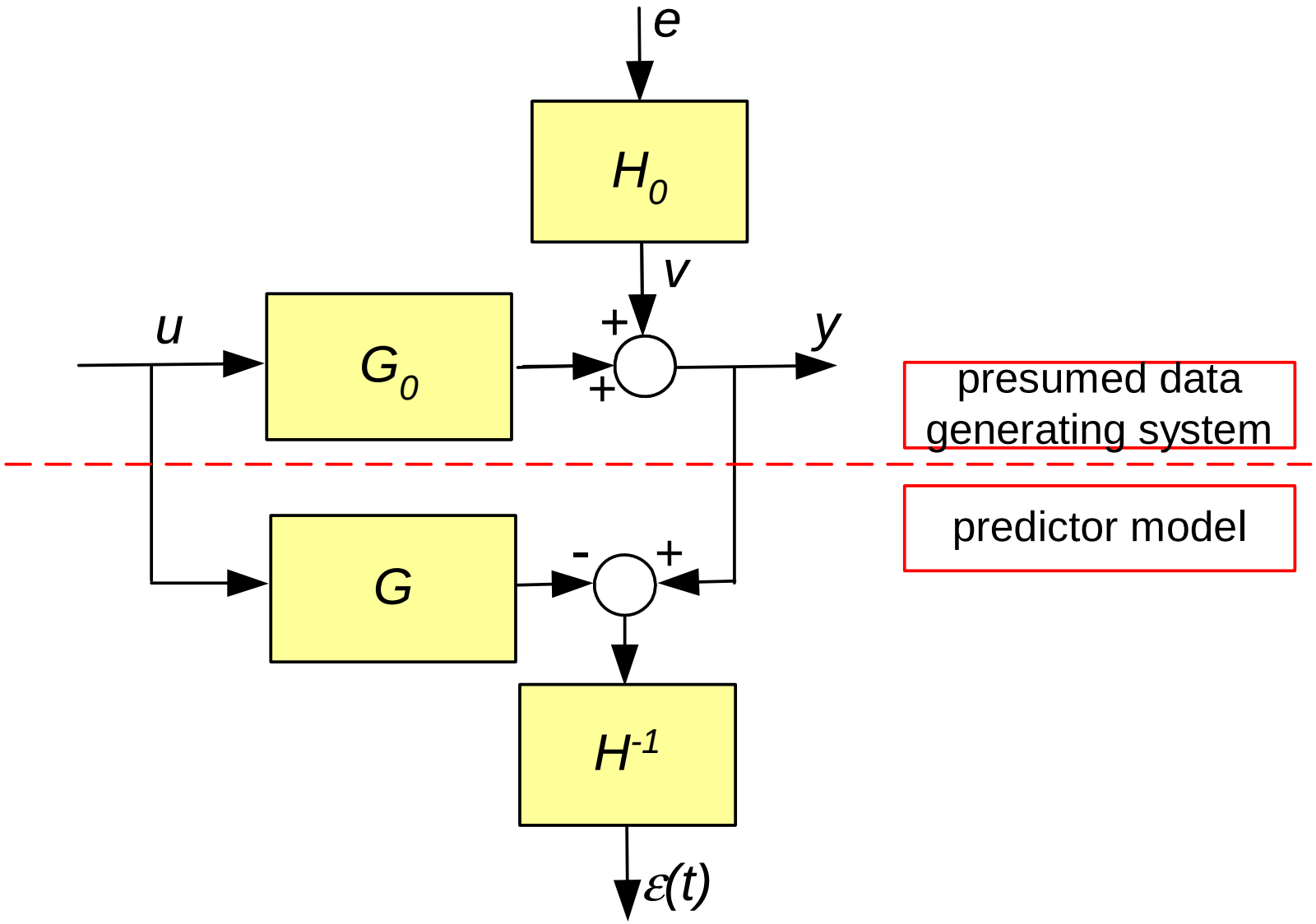
If the data has been generated by (G_0, H_0), then the predictor based on (G, H) leads to

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

the one-step-ahead prediction error.

$$\varepsilon(t) = H(q)^{-1}[y(t) - G(q)u(t)]$$

$\varepsilon(t)$ is going to serve as a basis for the identification criterion.


e
H_0
v
u
G_0
+
+
y
presumed data generating system
predictor model
G
-
+
H^{-1}
$\varepsilon(t)$

Prediction error:

$$\varepsilon(t) := y(t) - \hat{y}(t|t-1) = H^{-1}(q)[y(t) - G(q)u(t)]$$

where

- G, H reflects the hypothesized model, and
- y, u is data originating from a data-generating system.

The prediction error now becomes a quality tag for a model, in relation to a measured data sequence $\{u(t), y(t)\}_{t=1,\cdots N}$.
It compares the real measured output of a data generating system to the predicted output of a candidate model.
In the best case, $\varepsilon(t) = e(t)$ is white.

Is it important to include the noise model $H(\theta)$?

- It enables to make $\varepsilon(t)$ white
  (i.e. extract all information out of it)
- It will improve the statistical properties of the estimated model (see later)

Note that H describes the (frequency-dependent) properties of v, when driven by white noise.

Black box model structures and model sets

Predictor model: $(G(q), H(q))$

Model set:

$$\mathcal{M} = \{(G(q,\theta), H(q,\theta)), \theta \in \Theta \subset \mathbb{R}^d\}$$

A parametrization is used to represent models by real-valued coefficients

Example of parametrization:

- Coefficients in polynomial fractions $\Leftarrow \theta = (b_1 \ a_1 \ c_1 \ d_1)^T$

$$G(q,\theta) = \frac{b_1 q^{-1}}{1 + a_1 q^{-1}}; \quad H(q,\theta) = \frac{1 + c_1 q^{-1}}{1 + d_1 q^{-1}}$$

- Coefficients in series expansions
- Matrix coefficients in state space models

ARX Model structure

$$G(q,\theta) = \frac{B(q^{-1},\theta)}{A(q^{-1},\theta)}; \qquad H(q,\theta) = \frac{1}{A(q^{-1},\theta)}$$

with

$$\begin{aligned} B(q^{-1},\theta) &= q^{-n_k}\{b_0 + b_1 q^{-1} + \cdots + b_{n_b-1} q^{-n_b+1}\} \\ A(q^{-1},\theta) &= 1 + a_1 q^{-1} + \cdots + a_{n_a} q^{-n_a} \\ \theta &= [a_1 \ a_2 \ \cdots a_{n_a} \ b_0 \ b_1 \ \cdots \ b_{n_b-1}]^T. \end{aligned}$$

n_a, n_b are the number of parameters in the A and B polynomial.
n_k number of time delays

Predictor:

$$\hat{y}(t|t-1;\theta) = B(q^{-1},\theta)u(t) + [1 - A(q^{-1},\theta)]y(t)$$

Model structures

Model structure	$G(q,\theta)$	$H(q,\theta)$
ARX	$\frac{B(q^{-1},\theta)}{A(q^{-1},\theta)}$	$\frac{1}{A(q^{-1},\theta)}$
ARMAX	$\frac{B(q^{-1},\theta)}{A(q^{-1},\theta)}$	$\frac{C(q^{-1},\theta)}{A(q^{-1},\theta)}$
OE - Output Error	$\frac{B(q^{-1},\theta)}{F(q^{-1},\theta)}$	1
FIR	$B(q^{-1},\theta)$	1
BJ - Box-Jenkins	$\frac{B(q^{-1},\theta)}{F(q^{-1},\theta)}$	$\frac{C(q^{-1},\theta)}{D(q^{-1},\theta)}$

Properties of model structures

- Linearity-in-the-parameters (ARX)

$$\begin{aligned}\hat{y}(t|t-1;\theta) &= B(q^{-1})u(t) + (1 - A(q^{-1}))y(t) \\ &= \phi^T(t)\theta\end{aligned}$$

  is a linear function in θ.
  $\Rightarrow$ Important computational advantages.

- Independent parametrization of $G(q,\theta)$ en $H(q,\theta)$
  There are no common parameters in G and H.
  $\Rightarrow$ Advantages for independent identification of G and H.

Both properties to be utilized later on.

System in the model set

Data-generating system $\mathcal{S}$: $[G_0, H_0]$
Model set: $\mathcal{M}$: $\{[G(q,\theta), H(q,\theta)], \theta \in \Theta \subset \mathbb{R}^d\}$.

$$\mathcal{S} \in \mathcal{M}$$

denotes that the data generating system can exactly be represented within $\mathcal{M}$, i.e. $\exists \theta_0 \in \Theta$ such that

$$\begin{aligned} G(q,\theta_0) &= G_0(q) \\ H(q,\theta_0) &= H_0(q) \end{aligned}$$

The notion $G_0 \in \mathcal{G}$ with $\mathcal{G} = \{G(q,\theta), \theta \in \Theta \subset \mathbb{R}^d\}$ denotes that only G_0 can exactly be represented, i.e.

$$G(q,\theta_0) = G_0(q)$$

Identification criterion

Identification criterion

Consider the data-generating system:

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

and the parametrized model $G(q,\theta), H(q,\theta)$.
Denote: $\bar{V}(\theta) = \bar{\mathbb{E}}\varepsilon^2(t,\theta)$.
Then: $\bar{V}(\theta) \geq \sigma_e^2$, with equality for $\hat{\theta}$ if

$$\begin{aligned} G(q,\hat{\theta}) &= G_0(q) \\ H(q,\hat{\theta}) &= H_0(q) \end{aligned}$$

Uniqueness of this solution requires some more conditions, to be specified later.

Reasoning:

$$\varepsilon(t,\theta) = \frac{G_0(q) - G(q,\theta)}{H(q,\theta)} u(t) + \frac{H_0(q)}{H(q,\theta)} e(t)$$

u and e uncorrelated;
First term becomes 0 for $G(q,\hat{\theta}) = G_0(q)$.

$$\frac{H_0(q)}{H(q,\theta)} e(t) = [1 + \gamma_1(\theta) q^{-1} + \gamma_2(\theta) q^{-2} + \cdots] e(t)$$

$$\bar{\mathbb{E}} \varepsilon^2(t,\theta) = [1 + \gamma_1(\theta)^2 + \gamma_2(\theta)^2 \cdots] \sigma_e^2$$

is minimal for $\gamma_1(\hat{\theta}) = \gamma_2(\hat{\theta}) = \cdots = 0$.

$\Rightarrow$ Cost function minimum σ_e^2 is achieved for

$$G(q,\hat{\theta}) = G_0 \quad \text{and} \quad H(q,\hat{\theta}) = H_0.$$

Identification criterion

Power of prediction error is estimated from a data sequence through the quadratic criterion:

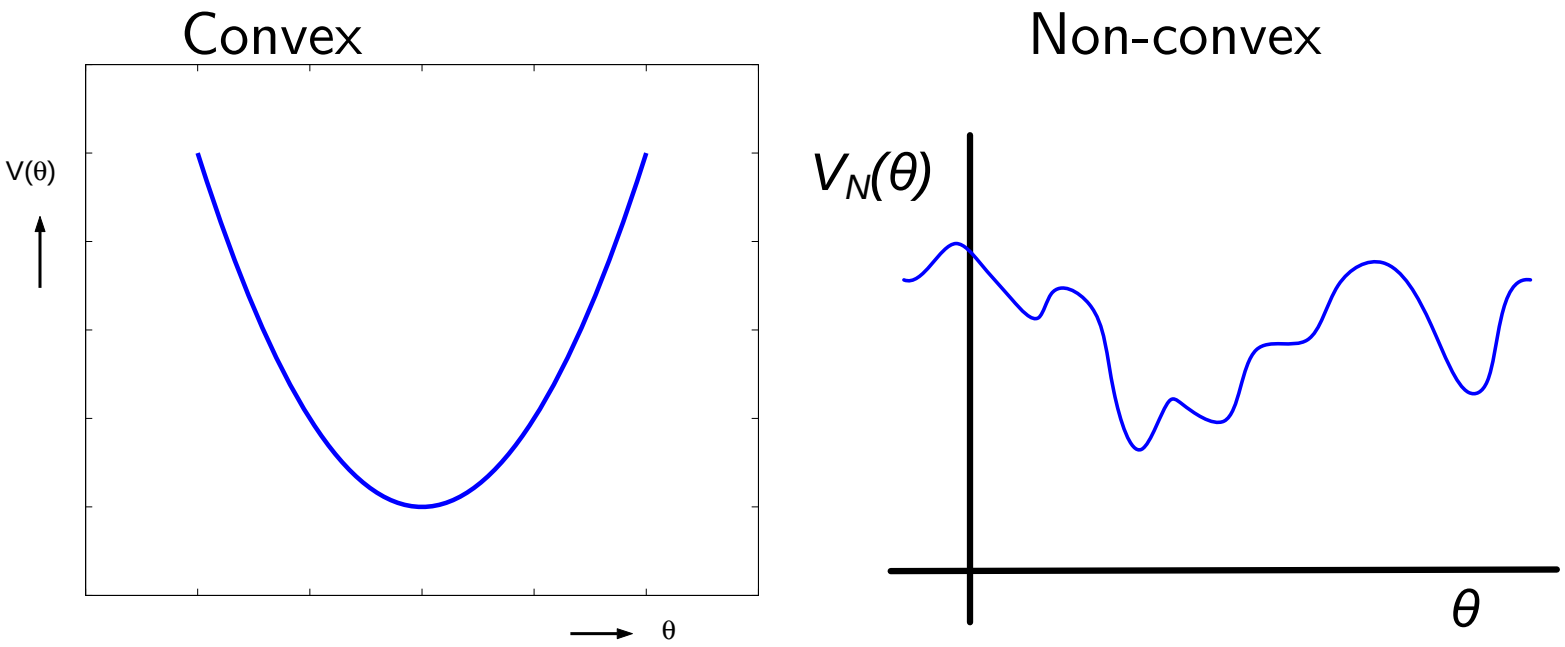
$$V_N(\theta, Z^N) = \frac{1}{N} \sum_{t=1}^{N} \varepsilon^2(t, \theta)$$

Parameter estimation through minimizing V_N:

$$\hat{\theta}_N = \arg\min_{\theta} V_N(\theta, Z^N)$$

Optimization problem that is convex or not....

Optimization:



Optimization is convex if the criterion is quadratic in θ.
In that situation, the solution can be found analytically.

Example

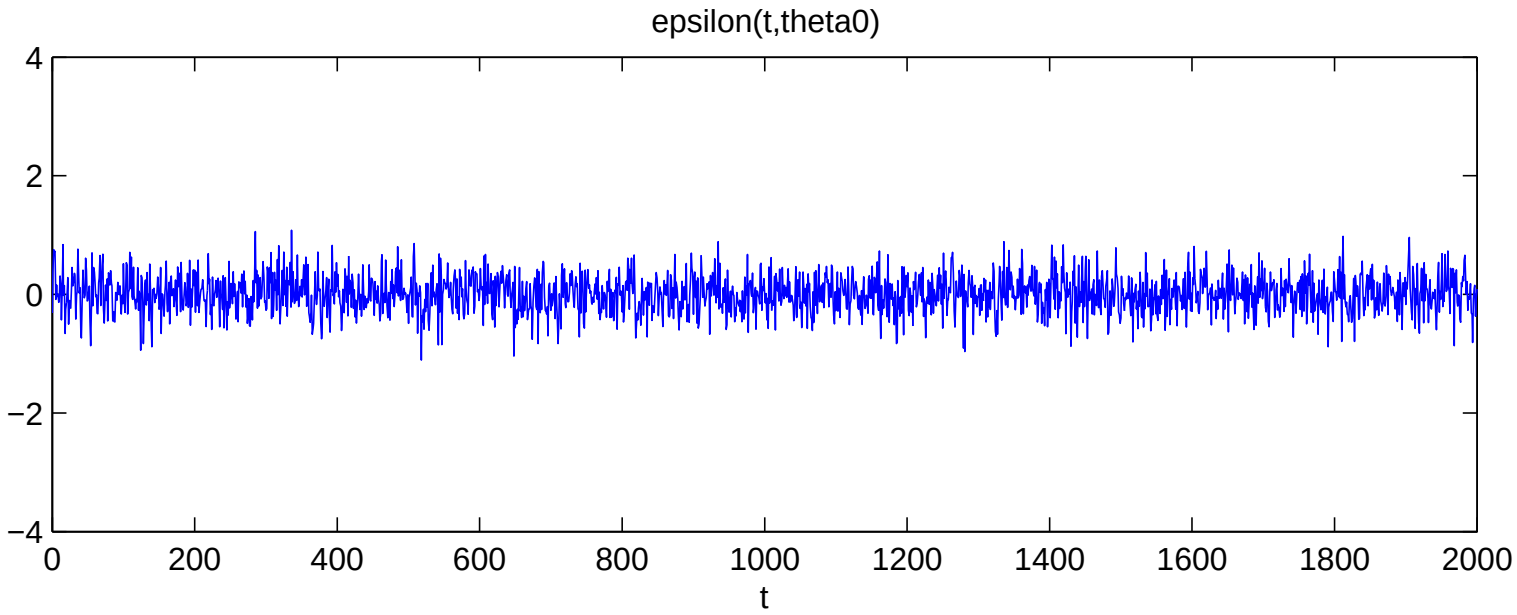
We have collected $N = 2000$ data $u(t)$ and $y(t)$ from the following data generating system

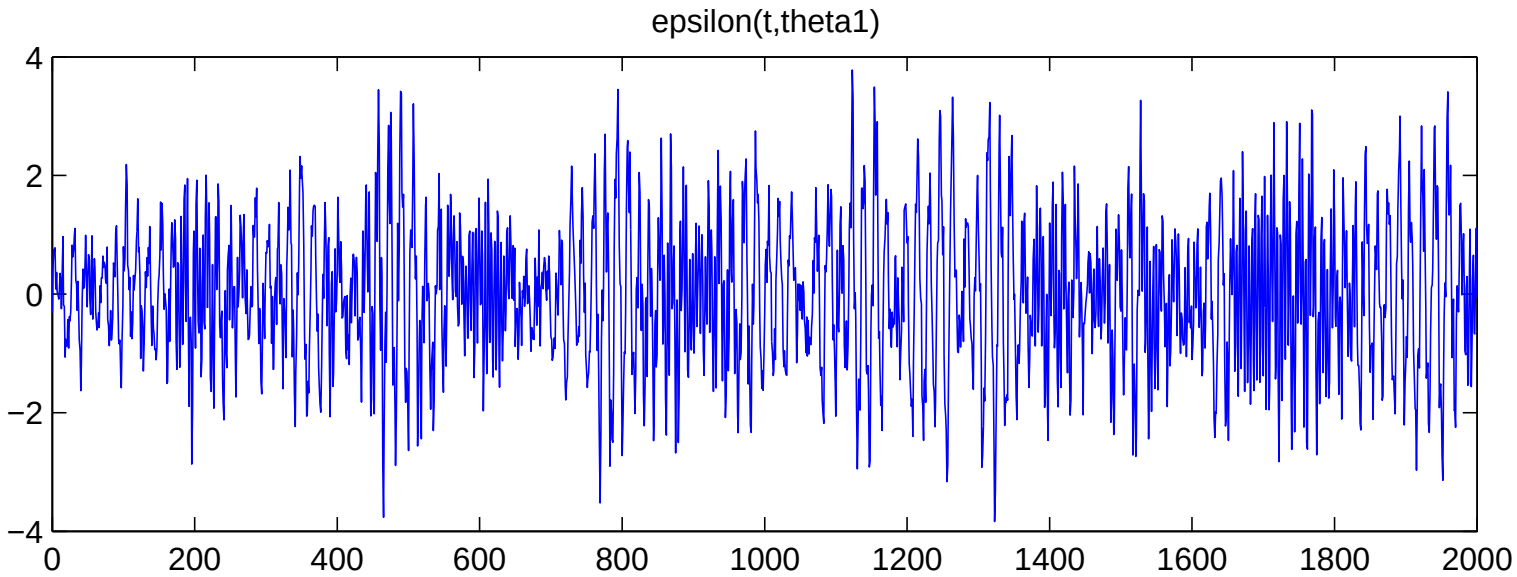
$$y(t) = G_0(q)u(t) + e(t)$$

with

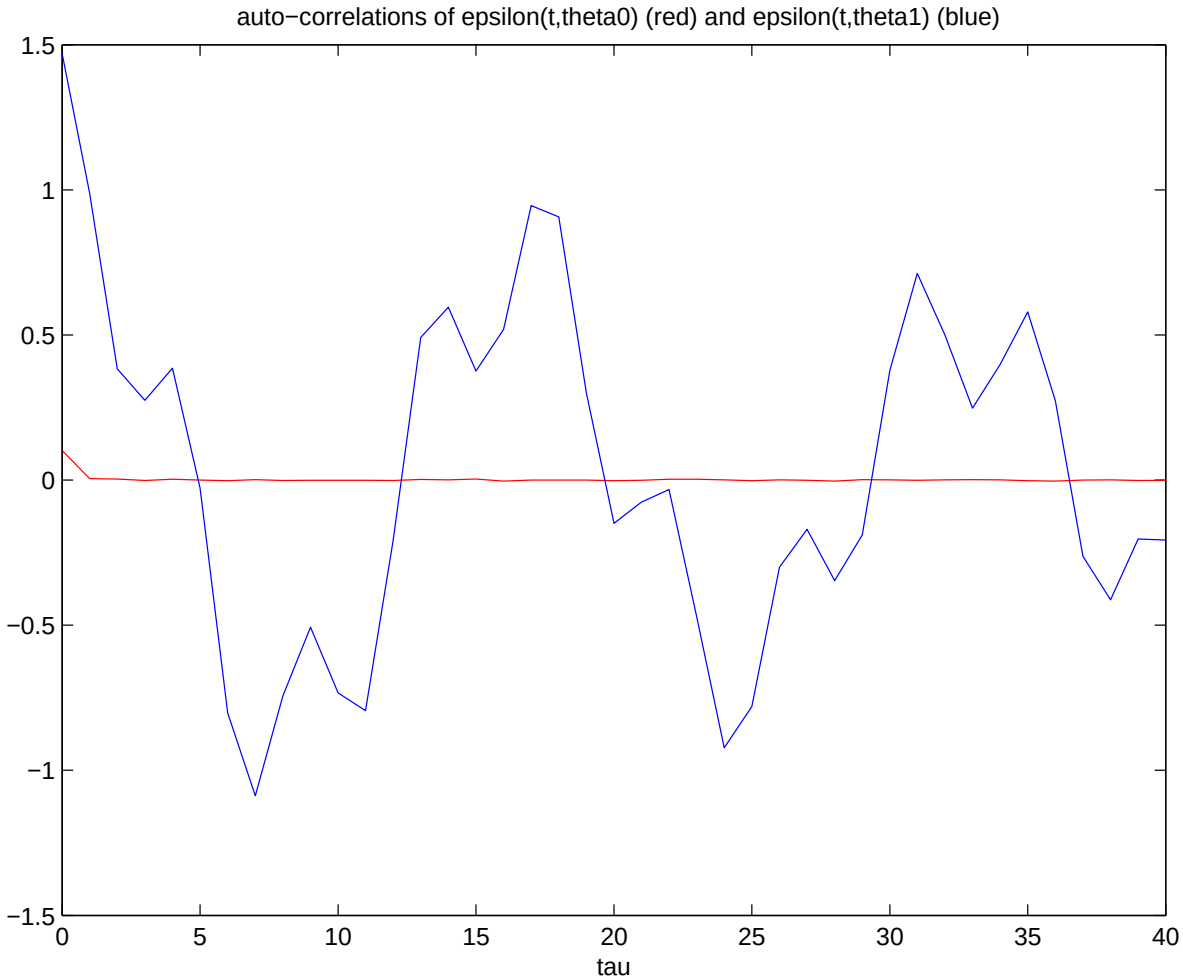
$$G_0(q) = \frac{q^{-3}\left(0.103 + 0.181q^{-1}\right)}{1 - 1.991q^{-1} + 2.203q^{-2} - 1.841q^{-3} + 0.894q^{-4}}.$$

We have computed $\varepsilon(t, \theta)$ $(t = 1...N)$ for $\theta = \theta_0$ and for another θ i.e. $\theta_1 \neq \theta_0$.


epsilon(t,theta0)
4
2
0
−2
−4
0
200
400
600
800
1000
1200
1400
1600
1800
2000
t


epsilon(t,theta1)
4
2
0
−2
−4
0
200
400
600
800
1000
1200
1400
1600
1800
2000

The estimated autocorrelation function $\hat{R}_\varepsilon^N(\tau)$ shows that $\varepsilon(t, \theta_0)$ well approaches the properties of a white noise as opposed to $\varepsilon(t, \theta_1)$.



Estimated power of $\epsilon(t, \theta_0)$: 0.1015 $(\sigma_e^2 = 0.1)$
Estimated power of $\epsilon(t, \theta_1)$: 1.4678
Note: the estimated power is $\hat{R}_\epsilon^N(0)$

Summary

- A one-step-ahead predictor has been defined to characterize how well an hypothesized model matches with a given data set
- A model structure defines how parameters enter the model $(G(q,\theta), H(q,\theta))$ that generates the predictor
- Model structures differ in the fact whether G and H are parametrized (in)dependently, and whether the predictor is linear in the parameters
- A (mostly quadratic) scalar-valued prediction error criterion measures the prediction error of a particular model, and can be used for estimating the parameters