System Identification
Lecture 3
PEM Analysis and Additional Topics

Paul Van den Hof

Control Systems Group
Eindhoven University of Technology

disc
dutch institute
of systems
and control

TU/e Technische Universiteit Eindhoven University of Technology

Contents

CRLB

ML estimator

MIMO models

Validation

AIC

Approximate modelling

Basis functions

Schedule

Apr 9	PVdH	Introduction; concepts
Apr 16	PVdH	Prediction error identification
Apr 23	PVdH	PEM analysis and additional topics
Apr 30	GB	Bayesian estimation; machine learning
May 7	JS	Frequency domain identification
May 14	JS	Nonlinear identification
May 28	XB	Experiment design
June 4	PVdH	Closed-loops and dynamic networks

Central question in estimation theory:

Does there exist - in a specified situation - a lower bound for the variance of a parameter estimator?

Cramé-Rao lower bound (CRLB)

Consider observations from a random variable $\mathbf{y}$ with pdf $f_{\mathbf{y}}(y, \theta)$, where θ is the unknown parameter. Then for *any* unbiased estimator $\hat{\boldsymbol{\theta}}$ of the parameter θ, its covariance matrix satisfies the inequality

$$cov(\hat{\boldsymbol{\theta}}) \geq J^{-1}$$

with the Fisher Information Matrix:

$$J = \mathbb{E}\left\{ - \frac{\partial^2}{\partial \theta^2} \log f_{\mathbf{y}}(\mathbf{y}; \theta) \Big|_{\theta=\theta_0} \right\}$$

Remarks on CRLB

- CRLB requires knowledge of pdf $f_{\mathbf{y}}(\mathbf{y};\theta)$
- CRLB generally requires exact knowledge of θ_0
  Exception: Gaussian pdf's with linear regression model
- It provides a lower bound for unbiased estimators
- Independent of the particular estimation method
- Useful for issues as
  - analysis, and
  - experiment design
- Estimator that reaches CRLB does not necessarily exist!

Maximum Likelihood Estimator

General estimator principle for situations where the probability density function (pdf) of the observed/measured variables is known.

Goal: Estimate an unknown parameter θ in the pdf of a random variable $\mathbf{y}$ on the basis of observations of y.

Example: random variable $\mathbf{y}$ has a Gaussian (unit variance) pdf with unknown $\mu_{\mathbf{y}}$

$$f_{\mathbf{y}}(y;\theta) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(y-\theta)^2}{2}}$$

- For given θ this is a pdf
- For given y and unknown θ this is a deterministic function of $\theta \rightarrow$ likelihood function $L(\theta)$

Maximum likelihood principle:

For given observations **y**, determine θ such that $L(\theta)$ is maximum.

(Choose that pdf that - a posteriori - makes the observed value of the considered variable most likely)

For 1 observation y:

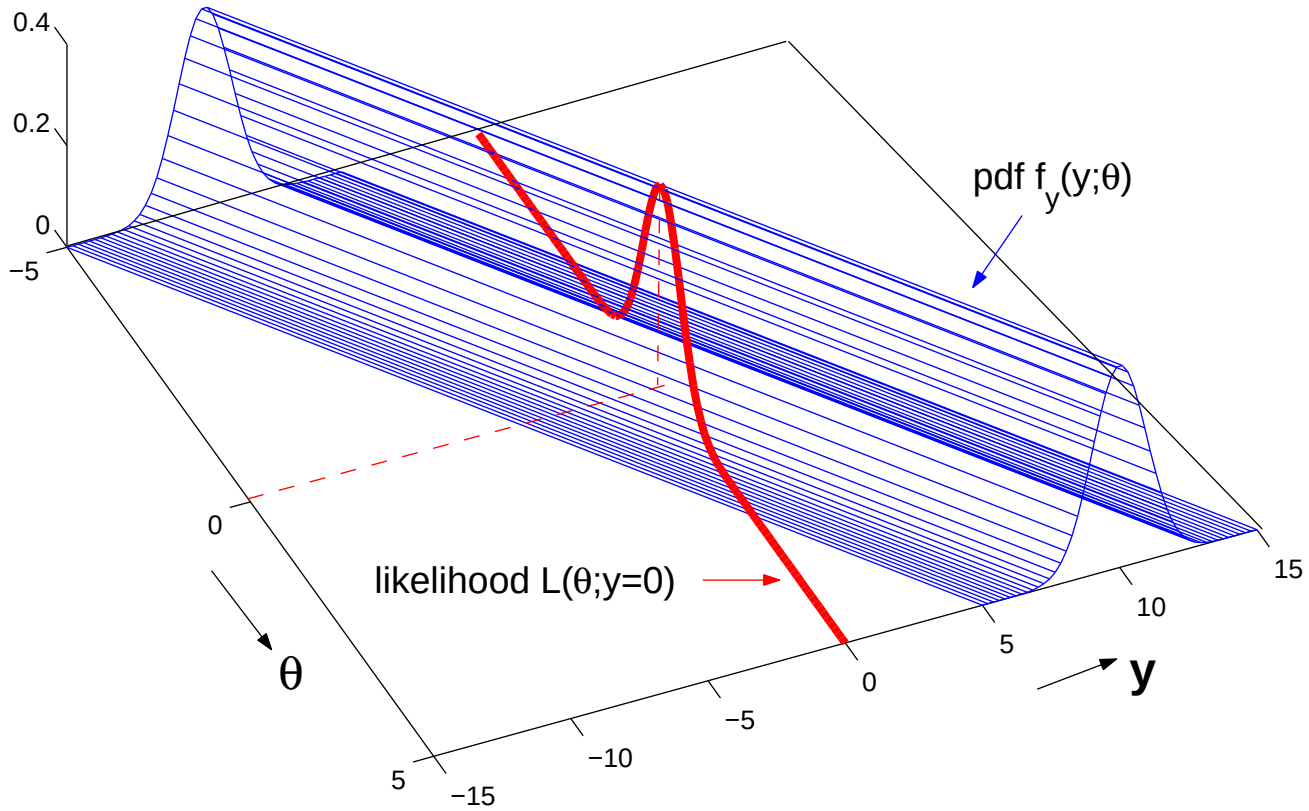
$$L(\theta) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(y-\theta)^2}{2}}$$

Maximizing $L(\theta)$ leads to: $\hat{\theta} = y$

For N independent observations y_i:

$$L(\theta) = f(y_1, \cdots, y_n; \theta) = \prod_{i=1}^{N} f(y_i; \theta)$$

Example: 1 observation with model $\mathbf{y} = \theta \cdot u + \mathbf{e}$



When observing $y = 0$ this leads to $\hat{\theta} = \arg\max_\theta L(\theta; y = 0)$.

Model: $\mathbf{y}(t) = \hat{y}(t|t-1;\theta) + \varepsilon(t)$, with $\varepsilon = \mathbf{e}$ for $\theta = \theta_0$.
If $\{\mathbf{e}(t)\}$ independent r.v.'s for different t with equal pdf f_e, then:

$$f_{\mathbf{y}}(x^N) = \prod_{t=1}^{N} f_{\mathbf{e}}(\mathbf{x}(t) - \hat{x}(t|t-1;\theta))$$

(probability density function of the observations)
and for the particular observation $x = y$ (a posteriori):

$$L_y(\theta; y^N) = \prod_{t=1}^{N} f_e(y(t) - \hat{y}(t|t-1;\theta))$$

the likelihood function, a deterministic function of θ.

$$L_y(\theta; y^N) = \prod_{t=1}^{N} f_e(\varepsilon(t,\theta)).$$

If f_e Gaussian:

$$L_y(\theta; y^N) = \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_e} e^{-\frac{\varepsilon(t,\theta)^2}{2\sigma_e^2}}$$

ML-estimator:

Maximizing L_y is equivalent to minimizing $-\log L_y$:

$$-\log L_y(\theta; y^N) = \frac{N}{2} log 2\pi + N \cdot log\sigma_e + \frac{1}{2\sigma_e^2} \sum_{t=1}^{N} \varepsilon(t,\theta)^2$$

If σ_e is either fixed or parametrized independent from θ, then

$$\min_{\theta} \sum_{t=1}^{N} \varepsilon(t,\theta)^2 \quad = \text{LS}$$

If the noise **e** on the data is a sequence of independent observations (white noise) from a Gaussian pdf with zero mean and equal variance for all observations, then the ML estimator is equal to the least squares (LS) prediction error estimator.

Or differently phrased:

> If in the PE setting, the noise disturbance e is Gaussian and $\theta^* = \theta_0$, then PE $=$ ML

Properties of the ML estimator

for $N \to \infty$:

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \to \mathcal{N}(0, NJ_N^{-1})$$

with J_N the Fisher Information Matrix, so that asymptotically in N

$$cov(\hat{\theta}_N) = J_N^{-1} \quad \text{(Cramér-Rao lower bound)}.$$

This implies that the ML estimator

- is consistent, provided that u is p.e. of sufficient order, and $\mathcal{S} \in \mathcal{M}$;
- asymptotically reaches the smallest possible variance (CRLB) over all unbiased estimators

no guarantees for properties in case of finite N

Multivariable models

What changes if we have multiple inputs and multiple outputs?

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

$y(t), e(t) \in \mathbb{R}^p$, $u(t) \in \mathbb{R}^m$, $cov(e) = \Lambda_0 \in \mathbb{R}^{p \times p}$

m: number of inputs

p: number of outputs

$$u(t) = \begin{bmatrix} u_1(t) \\ \vdots \\ u_m(t) \end{bmatrix}; \quad y(t) = \begin{bmatrix} y_1(t) \\ \vdots \\ y_p(t) \end{bmatrix}$$

- Scalar transfer function $\rightarrow$ transfer function matrix:

$$G(q) = \begin{bmatrix} G_{11}(q) & \cdots & G_{1m}(q) \\ \vdots & \vdots & \vdots \\ G_{p1}(q) & \cdots & G_{pm}(q) \end{bmatrix}$$

$$H(q) = \begin{bmatrix} H_{11}(q) & \cdots & H_{1p}(q) \\ \vdots & \vdots & \vdots \\ H_{p1}(q) & \cdots & H_{pp}(q) \end{bmatrix}$$

H stable, inversely stable, and monic, i.e. $\lim_{z\to\infty} H(z) = I$, or

$$H(z) = I + h_1 z^{-1} + h_2 z^{-2} + \cdots$$

- Predictor / prediction error remains the same:

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

$\varepsilon(t,\theta)$ is a p-dimensional vector;

- Identification criterion (scalar) is slightly generalized:

$$V_N(\theta) = \frac{1}{N}\sum_{t=0}^{N-1} \varepsilon^T(t,\theta)\Lambda^{-1}\varepsilon(t,\theta)$$

Λ^{-1} weighs the different components of ε with respect to eachother,
necessary e.g. in situations where the different components of $\varepsilon(t,\theta)$ have different scales (nm and km.....), or different noise levels.

- Model structures:
  This is the major change: different options for representing e.g. $G(q,\theta)$:
  - FIR:

$$G(q,\theta) = B_0 + B_1 q^{-1} + \cdots B_{n_b} q^{-n_b}, \qquad B_i \in \mathbb{R}^{p\times m}$$

  - Polynomial fractions, (ARX):

$$\begin{aligned} G(q,\theta) &= A(q,\theta)^{-1} B(q,\theta) \\ H(q,\theta) &= A(q,\theta)^{-1} \end{aligned}$$

    with

$$A(q,\theta) = I + A_1 q^{-1} + \cdots A_{n_a} q^{-n_a}$$

    possibly with $\{A_j\}_{j=1,n_a}$ restricted to being diagonal (i.e. same poles in each row of G)

Adaptations of the (SISO) theory:

- Maximum likelihood results for prediction error method remains valid provided that

$$\Lambda = \Lambda_0$$

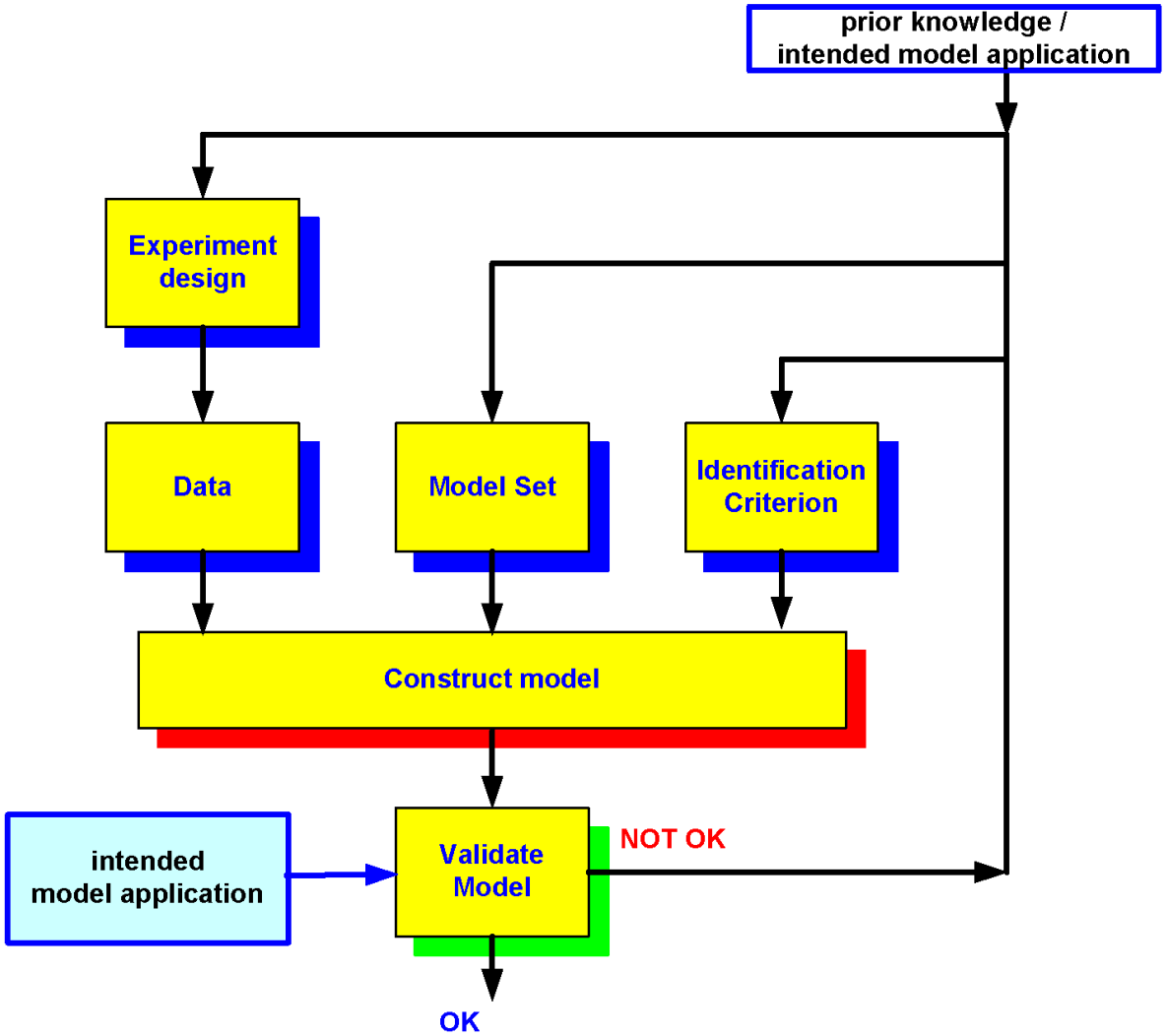
  i.e. the identification criterion is weighted with the inverse of the noise covariance matrix.
  When this is unknown it needs to be estimated $\rightarrow \Lambda(\theta)$, and the ML criterion gets more complicated.

- Persistence of excitation condition for the (multivariable) input signal u, becomes:

$$\Phi_u(\omega) > 0 \quad \text{(matrix inequality)}$$

  in a sufficient number of points ω.
  Interpretation: each input signal should have a component that is not linearly dynamically related to the other inputs.

Validation


prior knowledge / intended model application
Experiment design
Data
Model Set
Identification Criterion
Construct model
intended model application
Validate Model
NOT OK
OK

Residual-tests

$$\hat{R}_{\varepsilon}^{N}(\tau) := \frac{1}{N}\sum_{t=1}^{N-\tau}\varepsilon(t+\tau)\varepsilon(t) \qquad \hat{R}_{\varepsilon u}^{N}(\tau) := \frac{1}{N}\sum_{t=1}^{N-\tau}\varepsilon(t+\tau)u(t)$$

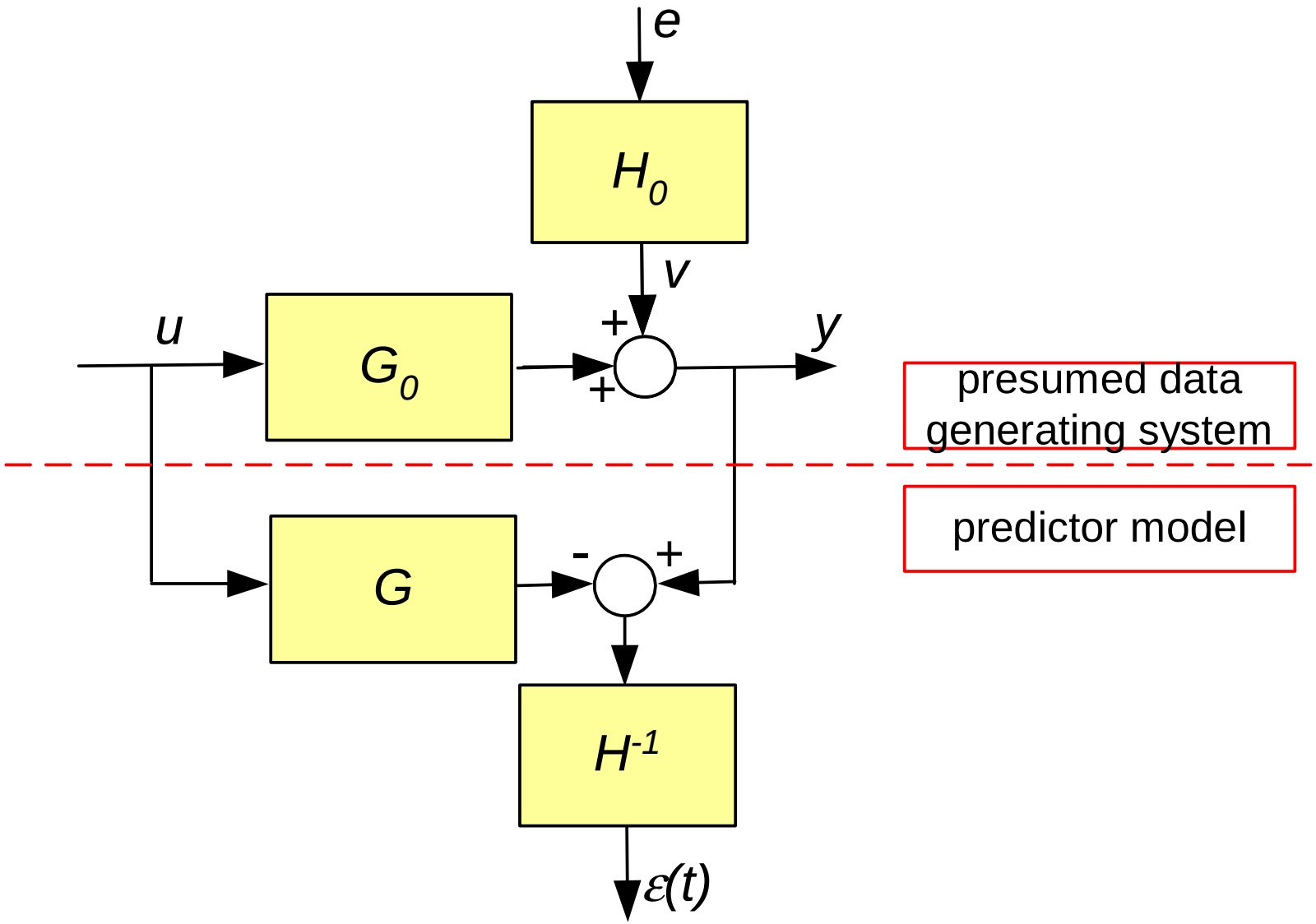
Test of hypotheses:

(a) $\varepsilon(t,\hat{\theta}_N)$ is a realization of a white noise process
(consistency $\hat{G}$, $\hat{H}$)
Evaluate

$$\hat{R}_{\varepsilon}^{N}(\tau) \to \delta(\tau)$$

(b) $\varepsilon(t,\hat{\theta}_N)$ is uncorrelated with past input samples
(consistency $\hat{G}$)
Evaluate

$$\hat{R}_{\varepsilon u}^{N}(\tau) \to 0, \qquad \tau \geq 0$$


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

Confidence intervals:

$$\sqrt{N}\frac{\hat{R}_\varepsilon^N(\tau)}{\hat{R}_\varepsilon^N(0)} \in As\ \mathcal{N}(0,1)$$

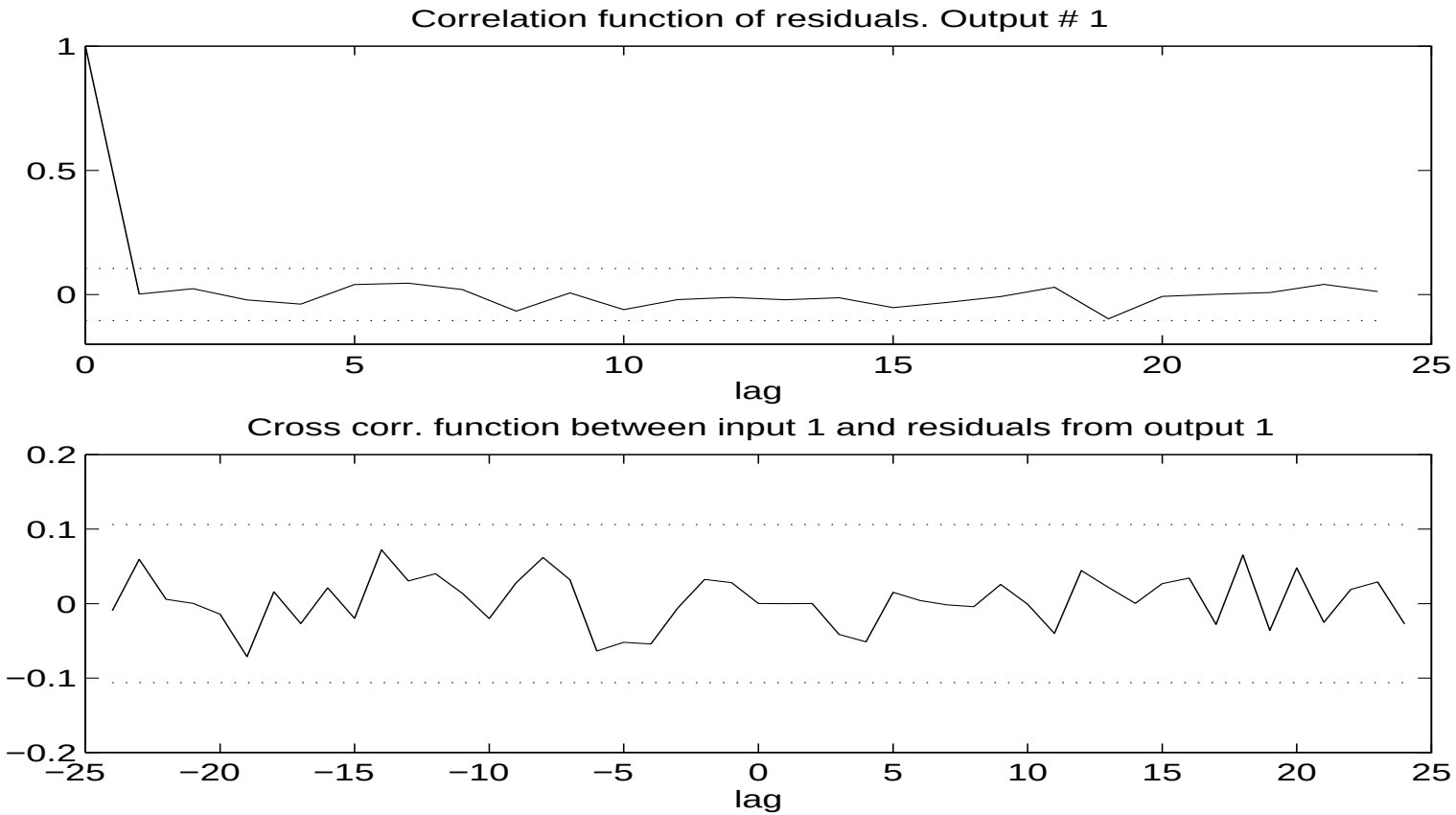
(valid for ε indeed being white noise)

$$\sqrt{N}\hat{R}_{\varepsilon u}^N(\tau) \in As\ \mathcal{N}(0,P)$$

$$P = \sum_{k=-\infty}^{\infty} R_\varepsilon(k)R_u(k)$$

(expressions for P based on the assumption that ε and u are indeed uncorrelated).

Through MATLAB-function RESID:



Confidence intervals of 99%.

The pointwise (over time) test should actually be replaced by a vector test on the full vector

$$\begin{bmatrix} \hat{R}^N_{\varepsilon u}(0) & \hat{R}^N_{\varepsilon u}(1) & \cdots & \hat{R}^N_{\varepsilon u}(n_\tau) \end{bmatrix}^T$$

taking account of the correlations between the several terms.

S. Douma, X. Bombois and P.M.J. Van den Hof (2008). Validity of the standard cross-correlation test for model structure validation. *Automatica, Vol. 44*, no. 5, pp. 1285-1294, 2008.

Cross-validation

For a given structure, compare

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

for $\hat{\theta}_N$ estimated for several model orders.
$\Rightarrow$ Large order leads to small V_N.

Avoiding of "overfit": Split the data: $Z^N = Z^{(1)}\ Z^{(2)}$
Estimate model on data $Z^{(1)}$:

$$\hat{\theta}_N^{(1)} = \arg\min_{\theta\in\Theta} V_{N^{(1)}}(\theta, Z^{(1)})$$

Evaluate criterion on data $Z^{(2)}$:

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

Akaike's Information Criterion (AIC)

Using cross-validation for comparing model sets

When distinguishing between estimation and validation data sets:

$\mathbb{E}V_N(\hat{\theta}_N)$: expected cost function on estimation data set

$\mathbb{E}\bar{V}(\hat{\theta}_N)$: expected cost function over validation data

taken over r.v. $\hat{\theta}_N$.

Then (Theorem 16.1, Ljung, 1999):

$$\mathbb{E}\bar{V}(\hat{\theta}_N) \approx \mathbb{E}V_N(\hat{\theta}_N) + \frac{1}{N} tr\, \bar{V}''(\theta^*) P_\theta$$

With $P_\theta = 2\sigma_e^2 \cdot [\bar{V}''(\theta_0)]^{-1}$ and $\theta^* = \theta_0$ it follows that

$$\mathbb{E}\bar{V}(\hat{\theta}_N) \approx V_N(\hat{\theta}_N) + \sigma_e^2 \frac{2n_\theta}{N} \quad \text{with } n_\theta = dim(\theta).$$

$$\mathbb{E}\bar{V}(\hat{\theta}_N) \approx V_N(\hat{\theta}_N) + \sigma_e^2 \frac{2n_\theta}{N} \quad \text{with } n_\theta = dim(\theta).$$

So, rather then minimizing $V_N(\theta)$ one should aim at minimizing

$$V_N(\theta) + \sigma_e^2 \frac{2n_\theta}{N}$$

i.e. extra penalty for "larger" model sets.

With σ_e^2 replaced by the estimate

$$\hat{\sigma}_e^2 = \frac{V_N(\hat{\theta}_N)}{1 - n_\theta/N}$$

the result is known as Akaike's Final Prediction Error criterion.

When applying the same reasoning to ML estimation with a Gaussian pdf, the resulting penalized cost function becomes

$$\log\left[\frac{1}{N}\sum_{t=1}^{N}\varepsilon^2(t,\theta)\right]+\frac{2n_\theta}{N}$$

which is known as Akaike's Information Criterion (AIC).

For further details see Ljung (1999), section 16.4.

Approximate Modelling

What can be said about

$$G(q, \theta^*) \quad H(q, \theta^*)$$

if $\mathcal{S} \notin \mathcal{M}$, and even $G_0 \notin \mathcal{G}$?

Can we characterize the approximative properties of identified models if they can not capture all dynamics of the data generating system?

We know (convergence result) that:

$$\theta^* = \arg\min_{\theta} \bar{V}(\theta)$$

and

$$\begin{aligned} \bar{V}(\theta) &= \bar{\mathbb{E}}\varepsilon(t,\theta)^2 \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi_\varepsilon(\omega)d\omega \end{aligned}$$

(Parsseval; both expressions are equal to $R_\varepsilon(0)$)

f-domain expression for the limit model

Combine:

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

with

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

then follows:

$$\varepsilon(t,\theta) = H(q,\theta)^{-1}\left[[G_0(q) - G(q,\theta)]u(t) + v(t)\right]$$

Consequence of the identification criterion:

$$\bar{V}(\theta) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi_\varepsilon(\omega)d\omega =$$

$$\bar{V}(\theta) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|G_0(e^{i\omega}) - G(e^{i\omega},\theta)|^2\Phi_u(\omega) + \Phi_v(\omega)}{|H(e^{i\omega},\theta)|^2}d\omega$$

This criterion plays the role of approximation criterion.

Alternative form

Write

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

then

$$\theta^* = \arg\min_{\theta}$$

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\frac{|G_0(e^{i\omega}) - G(e^{i\omega},\theta)|^2}{|H(e^{i\omega},\theta)|^2} \Phi_u(\omega) + \frac{|H_0(e^{i\omega}) - H(e^{i\omega},\theta)|^2}{|H(e^{i\omega},\theta)|^2} \sigma_e^2 \right] d\omega$$

Expression shows how θ^* is obtained.
Two mechanisms:

- Minimization of $\dfrac{|G_0(e^{i\omega}) - G(e^{i\omega},\theta)|^2 \Phi_u(\omega)}{|H(e^{i\omega},\theta)|^2}$
- Minimization of $\dfrac{|H_0(e^{i\omega}) - H(e^{i\omega},\theta)|^2 \sigma_e^2}{|H(e^{i\omega},\theta)|^2}$

Problems are coupled if $H(q,\theta)$ is parametrized.

Observation:
Type of approximation of G_0 by $G(q,\hat{\theta}_N)$ is dependent on noise model $H(q,\theta)$.

Special case: fixed noise model $H(q,\theta) = H_*(q)$ (e.g. OE).
Then

$$\theta^* = \arg\min_\theta \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|G_0(e^{i\omega}) - G(e^{i\omega},\theta)|^2 \Phi_u(\omega)}{|H_*(e^{i\omega})|^2} d\omega$$

(second term in integrand is θ-independent).
θ^* is determined by minimizing the integrated quadratic error $G_0 - G(\theta)$ with weighting function

$$\frac{\Phi_u(\omega)}{|H_*(e^{i\omega})|^2}$$

At those frequencies where the weighting function is large, the model error will be small.

Example

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

G_0 5th order; $\Phi_u(\omega) = 1$ (white noise).

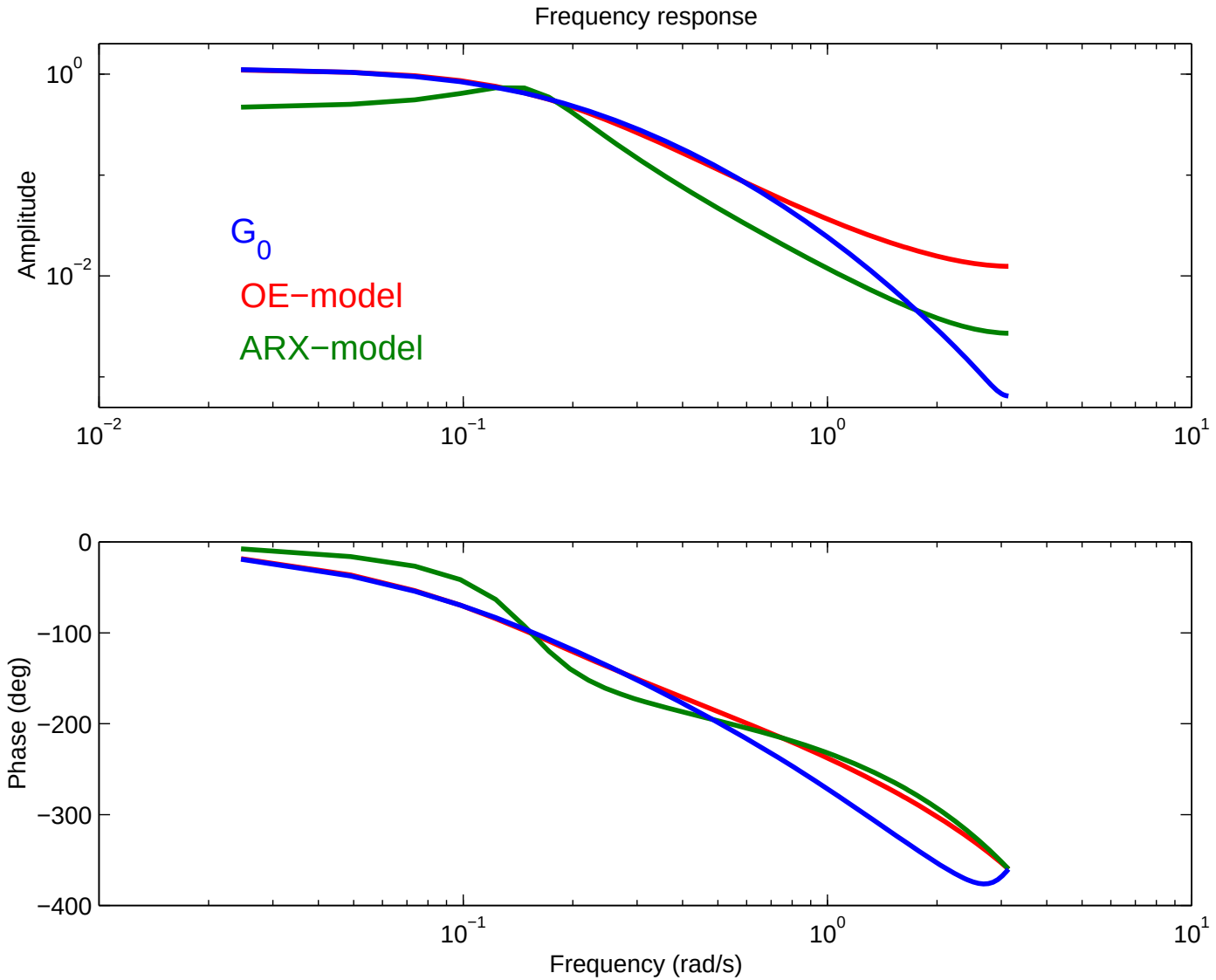
To illustrate the effect of approximation 2nd order models will be estimated:

- OE-model, 2nd order

$$y(t) = \frac{b_1 q^{-1} + b_2 q^{-2}}{1 + f_1 q^{-1} + f_2 q^{-2}} u(t) + e(t)$$

- ARX-model, 2nd order

$$(1 + a_1 q^{-1} + a_2 q^{-2}) y(t) = (b_1 q^{-1} + b_2 q^{-2}) u(t) + e(t)$$


Frequency response
Amplitude
10^0
10^{-2}
G_0
OE−model
ARX−model
10^{-2}
10^{-1}
10^0
10^1
Phase (deg)
0
−100
−200
−300
−400
10^{-1}
10^0
10^1
Frequency (rad/s)

Comparison of situations:

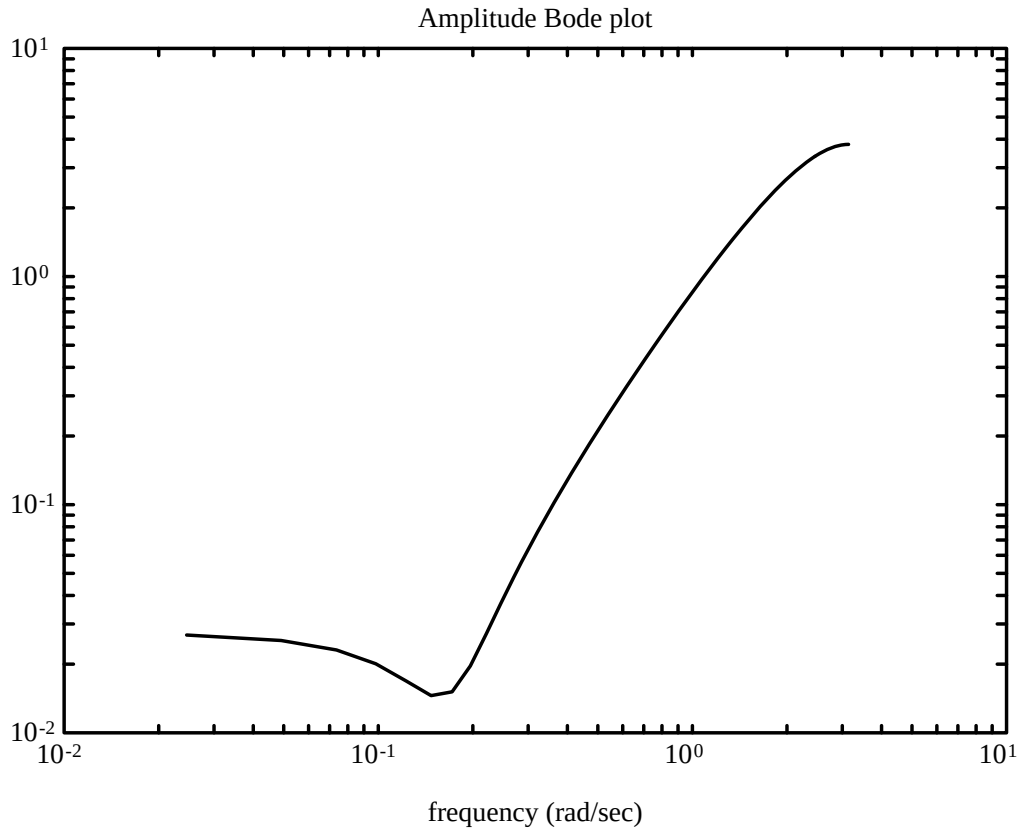
OE-model

$$\min \frac{1}{2\pi} \int_{-\pi}^{\pi} |G_0 - G(\theta)|^2 \Phi_u d\omega$$

ARX-model

$$\min \frac{1}{2\pi} \int_{-\pi}^{\pi} |G_0 - G(\theta)|^2 \Phi_u \cdot |A(e^{i\omega}, \theta)|^2 d\omega$$

Additional weighting with (a priori unknown) function.



Weighting function $|A(e^{i\omega}, \hat{\theta}_N)|$ in ARX-case.

General situation:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{|G_0(e^{i\omega})-G(e^{i\omega},\theta)|^2\Phi_u(\omega)+\Phi_v(\omega)}{|H(e^{i\omega},\theta)|^2}d\omega$$

In case of independent parametrization or fixed noise model $H = H_*$:
$G(q,\hat{\rho}_N)$ is obtained through

$$\hat{\rho}_N = \arg\min\frac{1}{2\pi}\int_{-\pi}^{\pi}|G_0-G(\rho)|^2\frac{\Phi_u}{|H_*|^2}d\omega$$

This holds for OE, BJ, FIR.

In other cases: compromise in which a.o.Φ_v is playing a role in the approximation of G_0.

Prefiltering of data

Prefiltering of the data with filter $L(q)$, leads to

$$\varepsilon_F(t,\theta) = L(q)\varepsilon(t,\theta)$$

and

$$\Phi_{\varepsilon_F}(\omega) = |L(e^{i\omega})|^2 \cdot \Phi_\varepsilon(\omega)$$

For a fixed noise model the weighting function becomes:

$$\frac{\Phi_u(\omega)|L(e^{i\omega})|^2}{|H_*(e^{i\omega})|^2}$$

Influencing the approximation criterion
(and so the resulting model)
by

- Choice of input spectrum Φ_u
- Choice of prefilter L
- Choice of noise model H

Example

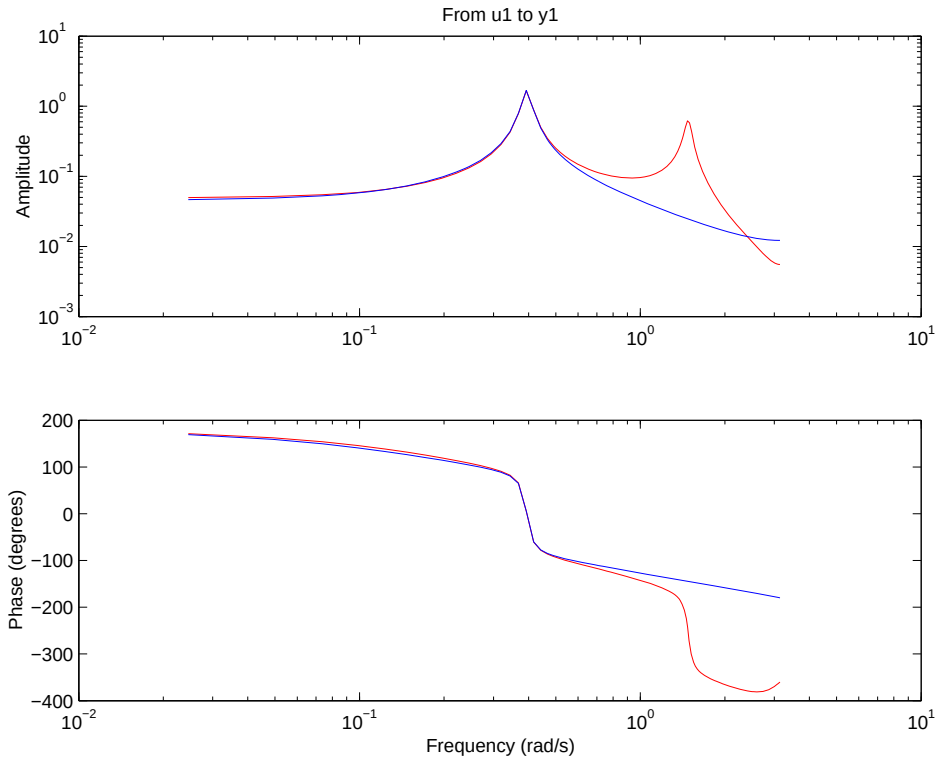
$$\mathcal{S}: \quad y(t) = G_0(q)u(t) + e(t)$$

with G_0 4th order with three delays.
We have to use a given set of data Z^N ($N = 5000$) for the identification where u is the sum of a white noise of variance 5 and four high-frequency sinusoids with amplitude 10, with frequencies: 1, 1.3, 1.5 and 2.0 rad/sec.
Objective: Using the given data, identify a good model $G(q, \hat{\theta}_N)$ for G_0 in the frequency range [0 0.7] in the reduced order model structure:

$$\mathcal{M} = OE(n_b = 2, n_f = 2, n_k = 3)$$

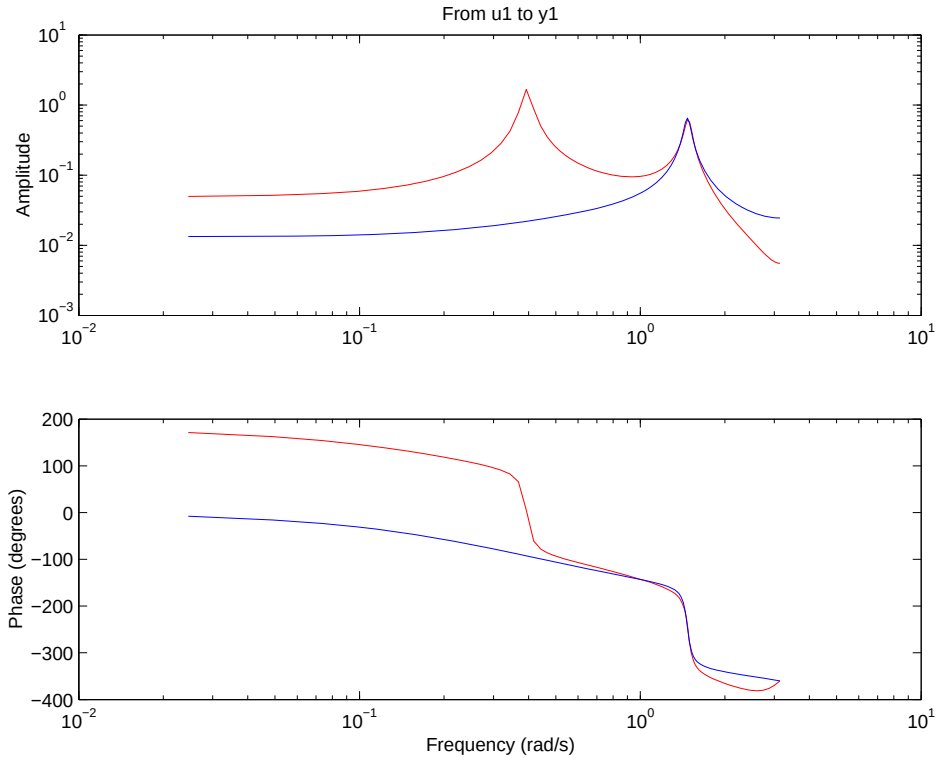
Since Z^N is given, the only degree of freedom we have is to use a pre-filter L to shape the bias error
We want a small bias error in the frequency range [0 0.7] $\Longrightarrow$ choose L such that $|L(e^{i\omega})|^2\Phi_u(\omega)$ is relatively (much) larger in the frequency range [0 0.7] than in [0.7 π]
$\Longrightarrow$ L Butterworth low pass filter of order 7 and cut-off frequency $0.7 rad/s$
We filter u and y collected from $\mathcal{S}$ by this L and we obtain filtered data with which we perform the identification in $\mathcal{M}$

G_0 (red) and $G(\hat{\theta}_N)$ (blue) identified with the filtered data.



$\Longrightarrow$ $G(\hat{\theta}_N)$ is OK.

What if we do not use a pre-filter L ?
G_0 (red) and $G(\hat{\theta}_N)$ (blue) identified with the data in Z^N



$\Longrightarrow$ $G(\hat{\theta}_N)$ is KO

Attractive model structures: OBF's

In PE results (Chapter 5) attractive properties of model structures:

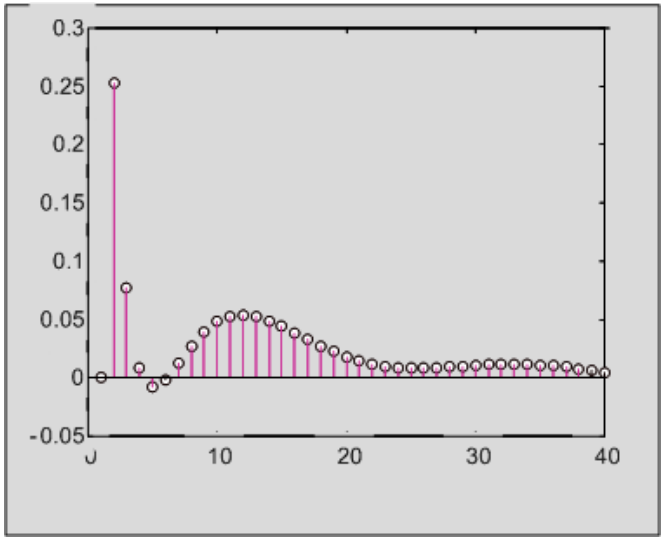
- Linearity-in-the-parameters: LS leads to convex optimization
- Output-Error structure: Consistent identification of G_0 irrespective of the noise model

Both properties are combined by the FIR-model:

$$\begin{aligned} G(z,\theta) &= \sum_{k=0}^{n_b} b_k z^{-k} \\ H(z,\theta) &= I \quad \text{(fixed)} \end{aligned}$$

Finite impulse response models:

$$G(z,\theta) = \sum_{k=0}^{n_b} b_k z^{-k}$$



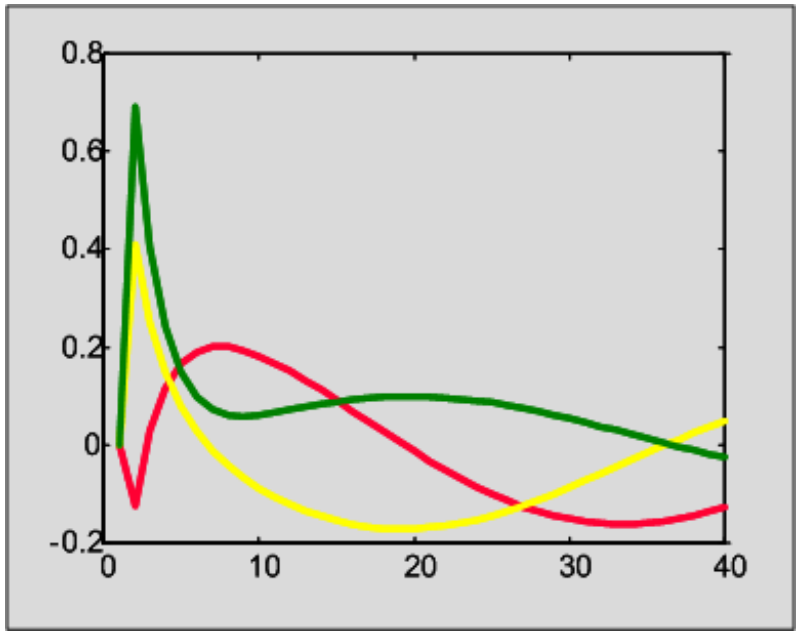
Disadvantage

For systems with relatively high sampling rates and small damping (long tails), the necessary number of parameters n_b can become very high

General aim towards parsimonious parametrizations:
Use parametrizations with few parameters, but in strategic locations

Generalized basis functions

From linear combinations of shifted pulse functions towards linear combinations of smartly chosen dynamic functions



$$G(z,\theta) = \sum_{k=0}^{n} c_k F_k(z)$$

Particular properties of the sequence $\{z^0, z^{-1}, z^{-2}, \cdots\}$:

- They constitute a complete basis for the systems space $\mathcal{H}_2$, Systems with a pulse response $g(k) \in \ell_2$ that satisfies $\sum_{k=0}^{\infty} |g(k)|^2 < \infty$.
- The inner product in this space is

$$< G_1, G_2 > := \frac{1}{2\pi} \int_{-\pi}^{\pi} G_1^*(e^{i\omega}) G_2(e^{i\omega}) d\omega = \sum_{k=0}^{\infty} g_1^*(k) g_2(k)$$

- All basis functions are orthonormal, i.e.

$$\begin{aligned} < z^{-i}, z^{-j} > &= 0, \quad i \neq j \\ &= 1, \quad i = j \end{aligned}$$

- When using a white noise input signal u, then

$$\mathbb{E}[q^{-i} u(t) \cdot q^{-j} u(t)] = 0 \quad i \neq j$$

Generalized orthonormal basis functions

- Laguerre functions
  Consider the sequence of functions

$$\{\frac{1}{z-a}, \frac{1}{(z-a)^2}, \frac{1}{(z-a)^3}, \cdots\}, \qquad (|a|<1)$$

  These functions constitute a complete basis for $\mathcal{H}_2$.
  After Gram-Schmidt orthogonalization:

$$F_k(z) = \frac{\sqrt{1-a^2}}{z-a}\left[\frac{1-az}{z-a}\right]^{k-1} \qquad k = 1, 2, \cdots$$

  with $|a| < 1$ real-valued pole.

Note: If $a = 0$ then the FIR structure results.

Generalization of this idea:

Consider the sequence of functions

$$\{\frac{1}{z-\xi_1}, \frac{1}{z-\xi_2}, \frac{1}{z-\xi_3}, \cdots\}, \qquad (|\xi_i| < 1)$$

Classical result (Walsh, 1935):

basis functions are complete in $\mathcal{H}_2$ if $\sum_{i=1}^{\infty}(1-|\xi_i|) = \infty$
(I.e. the poles should not tend to the unit circle).

After Gram-Schmidt orthogonalization:

$$F_k(z) = \frac{\sqrt{1-|\xi_k|^2}}{z-\xi_k} \underbrace{\prod_{i=1}^{k-1}\left[\frac{1-\xi_i^* z}{z-\xi_i}\right]}_{\text{all pass function}}$$

Again: For $\xi_i = 0$, $\forall i$, the FIR structure results.

Use in system identification

The considered model structure leads to

$$\hat{y}(t|t-1;\theta) = \sum_{k=1}^{n} c_k \cdot F_k(q)u(t)$$

Then $\varepsilon(t,\theta) = y(t) - \varphi^T(t)\theta$ with

$$\varphi(t) := [F_1(q)u(t)\ F_2(q)u(t)\ \cdots F_n(q)u(t)]^T.$$

The LS-estimate is simply obtained by

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

and all relevant properties of linear regression estimators are maintained.

Degrees of freedom: selection of basis poles

Pole selection usually done periodically:

$$\{\xi_1, \xi_2, \cdots \xi_n\} = \{\xi_1, \xi_2, \cdots \xi_{n_b}, \xi_1, \cdots \xi_{n_b}, \cdots\}.$$

Convergence result:

If $G_0(z)$ has poles $\{p_i\}_{i=1,..,n}$, and the basis is induced by a set of poles $\{\xi_j\}_{j=1,..,n_b}$, then the slowest eigenvalue that determines the convergence rate of

$$G_0(z) = \sum_{k=1}^{\infty} c_k F_k(z)$$

is given by

$$\max_i \prod_{j=1}^{n_b} \left| \frac{p_i - \xi_j^*}{1 - \xi_j p_i} \right|$$

$\Longrightarrow$ coefficients c_k decay quickly if pole sets are close

Example:

3rd order G_0: $p_{1,2} = 0.985 \pm 0.16i$; $p_3 = 0.75$

	ρ	$N_{0.01}$
FIR	0.9979	2191
Laguerre with $\xi_1 = 0.96$	0.9938	740
Kautz with $\xi_{1,2} = 0.97 \pm 0.1i$	0.9728	167
GOBF: $\xi_{1,2} = 0.97 \pm 0.1i$, $\xi_3 = 0.7$	0.9632	123
GOBF: $\xi_{1,2} = 0.98 \pm 0.15i$, $\xi_3 = 0.72$	0.7893	20

$N_{0.01}$:
number of coefficients before magnitude is decayed below 1%.

Features in identification with GOBF's

Use of "uncertain" prior knowledge in identification:

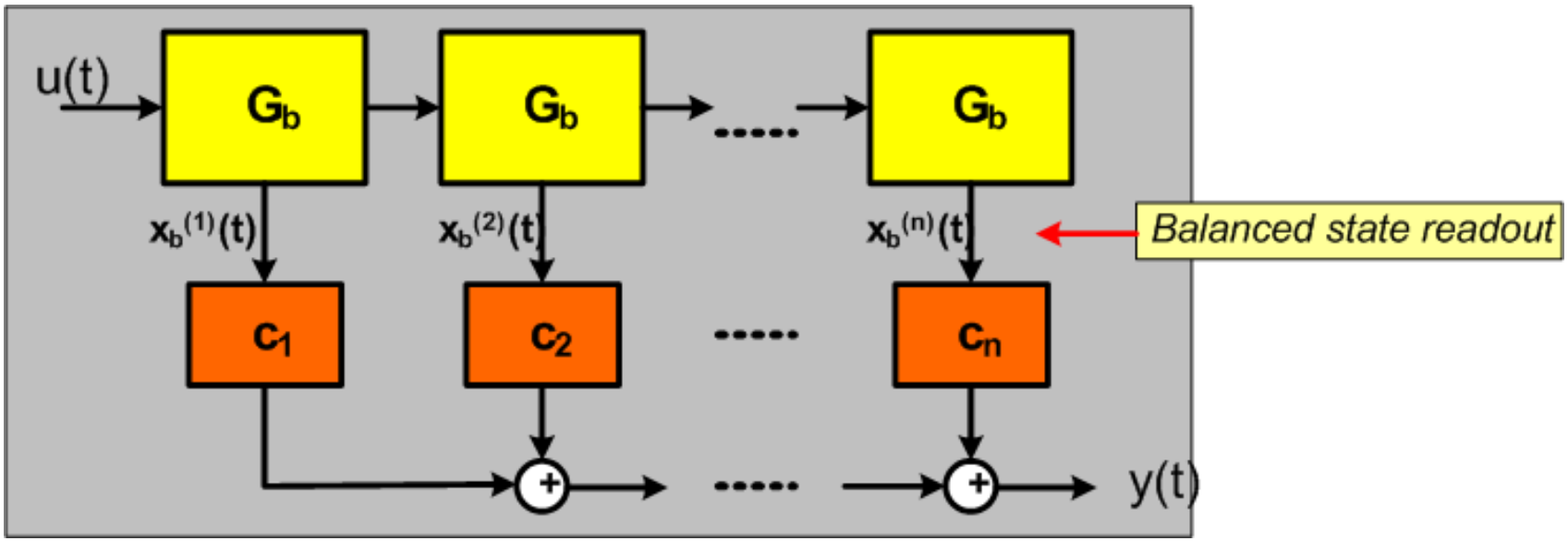
The more accurate the priors (pole information)
the more simple the ID-problem (less parameters)

but no loss of generality (for increasing n still all systems can be represented)

Additional properties:

- Variance analysis is available
- Noise model can be added (iteratively through GLS)
- Easily extendable to MIMO models
- Finite-time variance analysis
- Linear constraints can be added (static gain,...)

Alternative interpretation of basis construction



Through balanced-state representations of all-pass functions G_b generated by the poles $\{\xi_1, \cdots \xi_{n_b}\}$.
Generalization of so-called tapped delay line ($G_b = z^{-1}$).

Summary

- Interpretation of the PE identification criterion in a statistical ML framework, CRLB
- MIMO extensions
- Model validation and Akaike's criterion
- Approximate modeling results
- Extension of the classical model structures towards attractive GOBF's (Generalized Orthonormal Basis Functions)

Further reading in Heuberger, Van den Hof and Wahlberg (Springer 2005)