Data Driven Modeling

Lecture 1

Håkan Hjalmarsson

KTH - Royal Institute of Technology

hjalmars@kth.se

May 15, 2020

Outline

Outline

Introduction

- FEL3201 (8hp) / FEL3202 (12hp)
- Course elements
  - 13 lectures to provide an orientation
  - Q&A follow up the next lecture
  - Recommended reading in the form of lecture notes (continuously updated - feedback welcome!), and L. Ljung: system identification - Theory for the User (available online through KTHB)
  - Weekly homework problems. Peer correction.
  - Project. Groups of 2. Complete system id. problem. Preferrably real data. Optimal with something from your own research. Proposals due to hjalmars@kth.se by June 22. Deadline for reports September 15. 5 min. presentations. Date October TBD.
  - 48h take home exam starting at 9:00. Window: August 29 - September 13. Notify hjalmars@kth.se before August 25. Reminder at 8:30 at the day of the exam.

Introduction

- Course requirements
  - Homeworks: 80% solved
  - Exam: 50% for FEL3201. 65% for FEL3202.
  - Project: Approved report & presentation. Project for FEL3202 expected to be extensive (aim for conference paper).
- Many different areas blend together (Systems & Control theory, Mathematical statistics, Probability theory, Machine learning, Optimization theory,...)

Outline

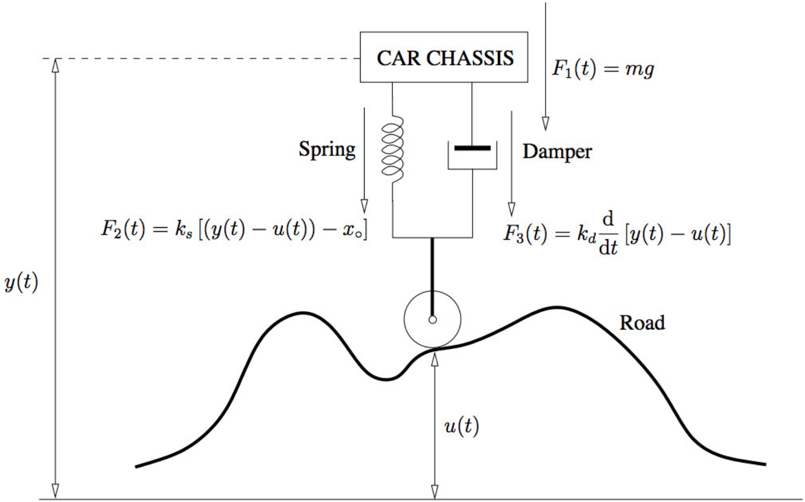
Course Outline

1. Introduction (Friday 15/5, 15-17) . Chapter 1-2 in Lecture Notes (LN). Chapter 1-2 in Ljung.
    - Signals and systems
    - The basic problem
    - Some examples
    - Introduction to parameter estimation
    - Some pitfalls
    - HW: 1.1 a-d (1.1f). 2.1 (2.2, 2.5)) Deadline Tuesday 26/5.
2. Probabilistic models (Tuesday 19/5, 10-12). Chapter 3 in LN. Chapter 4 in Ljung.
    - Models and model structures
    - Estimators
    - A probabilistic toolshed
3. Estimation theory and Wold decomposition (Tuesday 26/5, 10-12). Chapter 4 in LN. Chapter 3 in Ljung
    - Estimation theory
        - Information contents in random variables
        - Estimation of random variables
    - Wold decomposition
4. Unbiased parameter estimation (Friday 29/5, 15-17). Chapter 5 in LN. Chapter 7 in Ljung.
    - The Cramér-Rao lower bound
    - Efficient estimators
    - The maximum likelihood estimator
    - Data compression
    - Uniform minimum variance unbiased estimators
    - Best linear unbiased estimator (BLUE)
    - Using estimation for parameter estimation

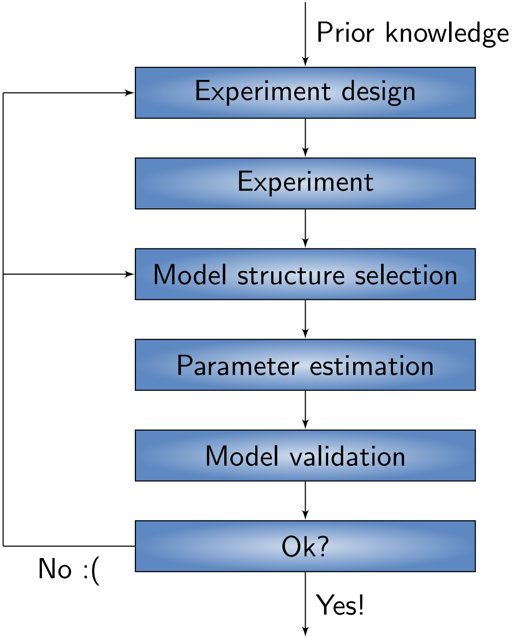
Course Outline

5. Biased parameter estimation (Tuesday 2/6, 10-12) . Chapter 6 in LN.
    - The bias-variance trade-off
    - The Cramér-Rao lower bound
    - Average risk minimization
    - Minimax estimation
    - Pointwise risk minimization
6. Asymptotic theory (Friday 5/6, 15-17). Chapter 7 in L.N. Chapter 8 in Ljung
    - Limits of random variables
    - Large sample properties of estimators
    - Using estimation for parameter estimation, part II
    - Large sample properties of biased estimators
7. Computational aspects (Tuesday 9/6, 08-10). Chapter 10 in Ljung.
    - Gradient based optimization
    - Convex relaxations
    - Integration by Markov Chain Monte Carlo (MCMC) methods
8. Case studies I (Friday 12/6, 10-12)
    - Parametric LTI models
    - Impulse response models
9. Case studies II (Tuesday 16/6, 10-12)
    - Uncertain input models
    - Nonlinear stochastic state-space models
10. Model accuracy (Friday 19/6, 15-17) Chapter 9 in Ljung.
11. Model structure selection and model validation (Tuesday 23/6, 10-12). Chapter 16 in Ljung
12. Experiment design (Tuesday 25/8, 10-12) . Chapter 13 in Ljung.
13. Continuous time identification (Friday 28/8, 15-17)

Introductory example: Shock absorber



System identification, an iterative procedure



Outline

Continuous time signals

Definition

The space $L_p(C)$, $0 < p < \infty$ *consists of all measurable functions* $F : C \to \mathbb{C}^{n\times m}$ *on* C *for which*

$$\|F\|_p := \left(\int_C \|F(t)\|_F^p dt \right)^{1/p} < \infty$$

The class $L_\infty(C)$ *consists of all measurable functions* $F : C \to \mathbb{C}^{n\times m}$ *on* C *for which*

$$\|F\|_\infty := \operatorname*{ess\,sup}_{t\in C} \overline{\sigma}(F(t)) < \infty$$

where $\overline{\sigma}(A)$ *denotes the largest singular value of the matrix* A.

The essential supremum for a real-valued function f is defined as

$$\operatorname*{ess\,sup}_{t\in C} f(t) = \inf\{a : \ f(t) \le a \text{ almost everywhere (a.e.) in } C$$

Continuous time signals

Fourier transform and its inverse

$$S(i\omega) = \int_{-\infty}^{\infty} s(t)e^{-i\omega t}dt, \quad \bar{s}(t) \qquad = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(i\omega)e^{i\omega t}d\omega$$

Theorem

i) *Suppose that $s \in L_1(\mathbb{R})$, then its Fourier transform S is uniformly continuous and vanishes at infinity.*

ii) *Suppose that $s \in L_1(\mathbb{R})$ and that its Fourier transform $S \in L_1(\mathbb{R})$.*

$$\text{Then } \bar{s}(t) = \int_{-\infty}^{\infty} S(i\omega)e^{i\omega t}d\omega$$

is continuous, vanishes at infinity and $\bar{s}(t) = s(t)$ a.e.

iii) *Suppose that $s \in L_p(\mathbb{R})$, $1 < p < \infty$, with Fourier transform S.*

$$\text{Then } \lim_{R\to\infty}\int_{|\omega|\le R} S(i\omega)e^{i\omega t}d\omega = s(t) \quad \text{a.e.}$$

Outline

Discrete time signals

Definition

The class ℓ_p, $0 < p < \infty$, *consists of all sequences* $\{s(t)\}$ *for which*

$$\|s\|_p := \left(\sum_k |s(t)|^p \right)^{1/p} < \infty$$

The class ℓ_∞ *consists of all sequences* $\{s(t)\}$ *for which*

$$\|s\|_\infty := \sup_t |s(t)| < \infty$$

$\ell_p \subset \ell_q$ for $1 \leq p < q \leq \infty$.
$s \in \ell_1 \Rightarrow$ Discrete Time Fourier transform (Fourier series)

$$S(e^{i\omega}) = \sum_{t=-\infty}^{\infty} s(t) e^{-i\omega t}$$

$$S \in L_1(\mathbb{T}), \Rightarrow \ \bar{s}(t) := \frac{1}{2\pi} \int_{-\pi}^{\pi} S(e^{i\omega}) e^{i\omega t} = s(t)$$

Discrete time signals

ℓ_2 and $L_2(\mathbb{T})$ Hilbert spaces with inner products

$$\langle s, v \rangle = \sum_t \mathrm{Trace}\{v^*(t)s(t)\}, \langle S, V \rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{Trace}\{V^*(e^{i\omega})S(e^{i\omega})\}$$

$b_k(\omega) = e^{i\omega k}$, complete orthonormal system for $L_2(\mathbb{T})$

Theorem

Any $S \in L_2(\mathbb{T})$ can be represented as $S(e^{i\omega}) = \sum_{t=-\infty}^{\infty} s(t)e^{-i\omega t}$ where

$$s(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S(e^{i\omega})e^{i\omega t}$$

What does $S = 0$ mean in $L_2(\mathbb{T})$? $\|S\|_2 = 0$. Equivalence classes.
ℓ_2 and $L_2(\mathbb{T})$ isomporphic: 1-1 relationship between elements.

Geometric properties preserved: $\langle S, V \rangle = \langle s, v \rangle$

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |S(e^{i\omega})|^2 d\omega = \|S\|_2^2 = \|s\|_2^2 = \sum_{t=-\infty}^{\infty} |s(t)|^2$$

Discrete time signals

z-transform: $S(z) := \sum_{k=-\infty}^{\infty} s(k)z^{-k}$ (Laurent series)
Holomorphic (analytic) in an annulus centered at the origin.

Definition

$H_p(\mathbb{T})$, $0 < p < \infty$ is the class of functions $F : \mathbb{T} \to \mathbb{C}^{n\times m}$ for which all elements are holomorphic in $|z| > 1$ and for which there is an $M < \infty$ such that

$$\int_{-\pi}^{\pi} \|F(re^{\omega})\|_F^p d\omega \leq M, \quad 1 < r < \infty$$

Theorem ($H_p(\mathbb{T})$ vs $L_p(\mathbb{T})$:)

Let $1 < p < \infty$. $S \in H_p(\mathbb{T}) \Leftrightarrow S(z) = \sum_{t=0}^{\infty} \bar{s}(t)z^{-t}$ where $\{\bar{s}(t)\}_{t=1}^{\infty}$ are the Fourier coefficiencts of some function in $L_p(\mathbb{T})$.

Dynamic systems

Linear time-invariant (LTI)

$$y(t) = \sum_{k=-\infty}^{\infty} g(k)u(t-k),$$

Short hand: $y(t) = G(q)u(t)$
where $G(q) = \sum_{k=-\infty}^{\infty} g(k)q^{-k}$ transfer function
z-transform: $Y(z) = G(z)U(z)$
Bounded-Input-Bounded-Output (BIBO) stability: $g \in \ell_1$
G maps signals to signals: e.g. $\ell_\infty \to \ell_\infty$. An operator

$$\|G\| = \sup_u \frac{\|Gu\|_\infty}{\|u\|_\infty} = \|g\|_1$$

$$\|G\| = \sup_u \frac{\|Gu\|_2}{\|u\|_2} = \sup_\omega |G(e^{i\omega})|$$

Dynamic systems

- Linear state space description

$$
\begin{aligned}
x(t+1) &= A(\theta)x(t) + B(\theta)u(t) + K(\theta)e(t) \\
y(t) &= C(\theta)x(t) + D(\theta)u(t) + e(t)
\end{aligned}
$$

  - $\{e(t)\}$ noise/disturbance
  - θ vector of unknown parameters
  - Black-box or (semi-)physical (grey-box)

- Non-linear

$$
\begin{aligned}
x(t+1) &= f(x(t), u(t), w(t), \theta) \\
y(t) &= h(x(t), u(t), e(t), \theta)
\end{aligned}
$$

Common linear black-box structures

- FIR

$$
\begin{aligned}
y(t) &= b_1 u(t-1) + \ldots b_n u(t-n) + e(t) \\
&= \begin{bmatrix} u(t-1) & \ldots & u(t-n) \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} + e(t) = \varphi^T(t)\theta + e(t)
\end{aligned}
$$

Compact form:
$y(t) = B(q)u(t) + e(t) = (b_1 q^{-1} + \ldots + b_n q^{-n})u(t) + e(t).$

- General:

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

where G and H are rational discrete-time transfer functions.

Common linear black-box structures

- FIR

$$y(t) = b_1 u(t-1) + \dots b_n u(t-n) + e(t)$$
$$= \begin{bmatrix} u(t-1) & \dots & u(t-n) \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} + e(t) = \varphi^T(t)\theta + e(t)$$

Compact form:
$y(t) = B(q)u(t) + e(t) = (b_1 q^{-1} + \dots + b_n q^{-n})u(t) + e(t).$

- General:

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

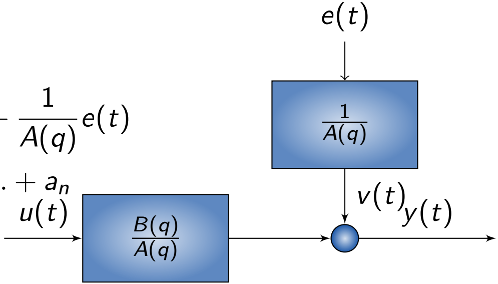
where G and H are rational discrete-time transfer functions.

Common linear black-box structures

- General: $y(t) = G(q,\theta)u(t) + H(q,\theta)e(t)$
- ARX

$$y(t) = \frac{B(q)}{A(q)}u(t) + \frac{1}{A(q)}e(t)$$

$$A(q) = 1 + a_1 + \ldots + a_n$$



Can be written $A(q)y(t) = B(q)u(t) + e(t)$
which is equivalent to

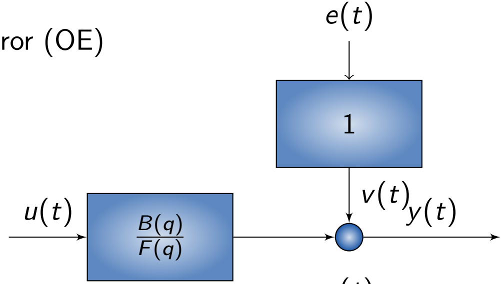
$$y(t) = \varphi^T\theta + e(t)$$

$$\varphi(t) = \begin{bmatrix} -y(t-1) & \ldots - y(t-n) & u(t-1) & \ldots & u(t-n) \end{bmatrix}^T$$

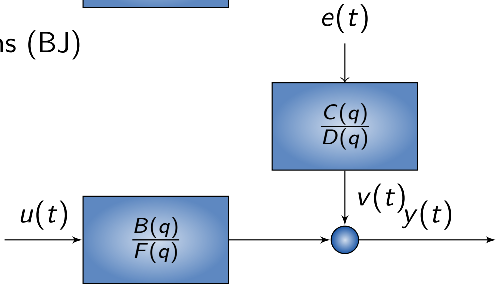
$$\theta = \begin{bmatrix} a_1 & \ldots & a_n & b_1 & \ldots & b_n \end{bmatrix}^T$$

Common linear black-box structures

- Output-Error (OE)



- Box-Jenkins (BJ)



Continuous time models

$$\dot{x}(t) = \mathcal{A}(\theta)x(t) + \mathcal{B}(\theta)u(t) + w(t)$$
$$y(t) = \mathcal{C}(\theta)x(t) + \mathcal{D}(\theta)u(t) + v(t)$$

Sampling gives

$$x(t+1) \approx A(\theta)x(t) + B(\theta)u(t) + K(\theta)e(t)$$
$$y(t) \approx C(\theta)x(t) + D(\theta)u(t) + e(t)$$

Important to use correct intersample behaviour of input.

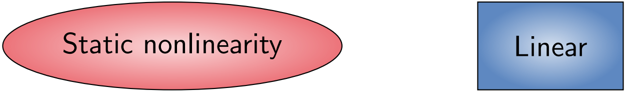
Common nonlinear black-box models

- Predictor models

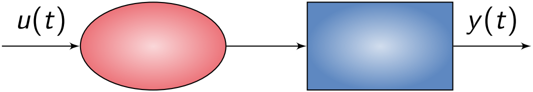
$$y(t) = g(\varphi(t), \theta) + e(t)$$

  where $\varphi(t)$ (nonlinear transformations of) past inputs and outputs.
  - Neural networks
  - Radial basis functions
  - NLARX: $\varphi(t)$ past inputs and outputs
  -
  -
  -
- Block oriented models

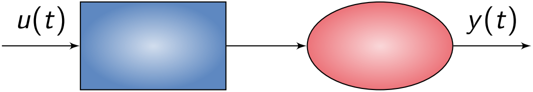
Block-oriented models



- Hammerstein (nonlinear actuator)



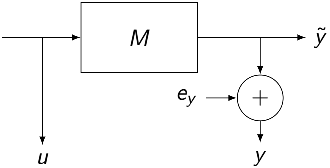
- Wiener (nonlinear sensor)



- Hammerstein-Wiener

Outline

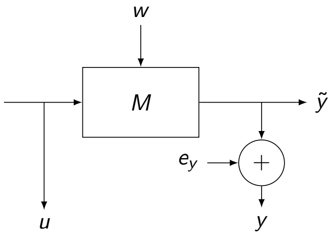
Example 1: Scalar LTI model



$$\mathbf{y} = \Phi \mathbf{g} + \mathbf{e}_y$$

- Measurements: $\mathbf{y} \in \mathbb{R}^N$ (u known exactly and can be considered part of the model)
- Unknowns: $\mathbf{g} \in \mathbb{R}^n$, $\mathbf{e}_y \in \mathbb{R}^N$

Example 2: Scalar LTI state-space model

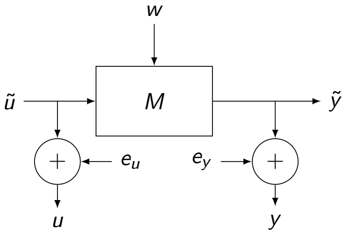


$$\mathbf{x} = F(\boldsymbol{\theta})\mathbf{u} + G(\boldsymbol{\theta})\mathbf{w}$$

$$\mathbf{y} = H(\boldsymbol{\theta})\mathbf{x} + \mathbf{e}_y, \quad \mathbf{y} \in \mathbb{R}^N$$

- Measurements: $\mathbf{y} \in \mathbb{R}^N$
- Unknowns: $\mathbf{w} \in \mathbb{R}^{mN}$, $\boldsymbol{\theta} \in \mathbb{R}^{m^2+2m}$, $\mathbf{e}_y \in \mathbb{R}^N$

Example 3: Scalar LTI state-space EIV model



$$\begin{aligned} \mathbf{x} &= F(\boldsymbol{\theta})\mathbf{u} + G(\boldsymbol{\theta})\mathbf{w} \\ \mathbf{u} &= \tilde{\mathbf{u}} + \mathbf{e}_u \\ \mathbf{y} &= H(\boldsymbol{\theta})\mathbf{x} + \mathbf{e}_y \end{aligned}$$

- Model order: m
- Measurements: $\mathbf{u} \in \mathbb{R}^N$, $\mathbf{y} \in \mathbb{R}^N$
- Unknowns: $\mathbf{w} \in \mathbb{R}^{mN}$, $\boldsymbol{\theta} \in \mathbb{R}^{m^2+2m}$, $\mathbf{e}_u \in \mathbb{R}^N$, $\mathbf{e}_y \in \mathbb{R}^N$

Outline

Key issue #1: More unknowns than measurements

Collect all unknowns in $\boldsymbol{\xi} \in \boldsymbol{\Xi}$.

- Model: $\mathbf{z}(\boldsymbol{\xi})$
- Data: $\mathbf{z}$

Unfalsified parameter set: $\boldsymbol{\Xi}(\mathbf{z}) := \{\boldsymbol{\xi} \in \boldsymbol{\Xi} : \ \mathbf{z}(\boldsymbol{\xi}) = \mathbf{z}\}$

Any further inference must be based on introducing a prejudice among the $\boldsymbol{\xi}$'s in $\boldsymbol{\Xi}(\mathbf{z})$. How can we do this? Ranking!

Introduce ranking function: $p(\boldsymbol{\xi}) \geq 0$, $\int_{\boldsymbol{\Xi}} p(\boldsymbol{\xi}) d\boldsymbol{\xi} = 1$

Maximum of rankings estimate:

$$\hat{\boldsymbol{\xi}}_M(\mathbf{z}) := \arg\max_{\boldsymbol{\xi} \in \boldsymbol{\Xi}(\mathbf{z})} p(\boldsymbol{\xi})$$

Notice that the ranking function has nothing to do with the data. The only connection to the data is that we maximize over the unknowns consistent with the data.

Encoding the set of unfalsified models

Recall Dirac's delta function: $\int f(t)\delta(t)dt = f(0)$
Multivariable version:

$$\boldsymbol{\delta}(\mathbf{x}) := \prod_{k=1}^{n} \delta(x(k)), \quad \mathbf{x} = \begin{bmatrix} x(1) & \dots & x(n) \end{bmatrix}^T \in \mathbb{R}^n$$

The joint ranking of model parameters $\boldsymbol{\xi}$ and observations $\mathbf{z}$:

$$p(\boldsymbol{\xi}, \mathbf{z}) := p(\boldsymbol{\xi})\boldsymbol{\delta}(\mathbf{z} - \mathbf{M}(\boldsymbol{\xi})),$$

Gives:

$$\hat{\boldsymbol{\xi}}(\mathbf{z}) = \arg\max_{\boldsymbol{\xi}} p(\boldsymbol{\xi}, \mathbf{z})$$

Key issue #1: More unknowns than measurements

Alternative: *Average of rankings estimate*:

$$\hat{\boldsymbol{\xi}}_A(\mathbf{z}) := \frac{\int_{\boldsymbol{\Xi}(\mathbf{z})} \boldsymbol{\xi}\ p(\boldsymbol{\xi})\ d\boldsymbol{\xi}}{p_z(\mathbf{z})} = \frac{\int \boldsymbol{\xi}\ p(\boldsymbol{\xi}, \mathbf{z})\ d\boldsymbol{\xi}}{p_z(\mathbf{z})}$$

$$\text{where } p_z(\mathbf{z}) := \int_{\boldsymbol{\Xi}(\mathbf{z})} p(\boldsymbol{\xi})\ d\boldsymbol{\xi} = \int p(\boldsymbol{\xi}, \mathbf{z})\ d\boldsymbol{\xi}$$

Simplification: Use $p(\boldsymbol{\xi}|\mathbf{z}) := p(\boldsymbol{\xi}, \mathbf{z})/p_z(\mathbf{z})$:

$$\hat{\boldsymbol{\xi}}_A(\mathbf{z}) = \int \boldsymbol{\xi}\ p(\boldsymbol{\xi}|\mathbf{z}) d\boldsymbol{\xi}$$

That's it folks - the course is finished!

From here on it can only become more confusing

Example 1 cont'd

$$\mathbf{y}(\mathbf{g}, \mathbf{e}_y) = \Phi\mathbf{g} + \mathbf{e}_y, \quad \boldsymbol{\xi} = \begin{bmatrix} \mathbf{g} \\ \mathbf{e}_y \end{bmatrix}$$

Introduce ranking:

$$p(\boldsymbol{\xi}) = \mathcal{N}(\mathbf{e}_y; 0, \lambda_{e_y} I)\, \mathcal{N}(\mathbf{g}, 0, K_g)$$

- Stochastic modeling is just a convoluted way to rank
- $p(\boldsymbol{\xi})$ pdf for all unknowns
- $p_y(\mathbf{y})$ pdf for $\mathbf{y}$

Estimates:

$$\hat{\boldsymbol{\xi}}_M(\mathbf{y}) := \arg\max_{\boldsymbol{\xi} \in \boldsymbol{\Xi}(\mathbf{y})} \mathcal{N}(\mathbf{e}_y; 0, \lambda_{e_y} I)\, \mathcal{N}(\mathbf{g}, 0, K_g) \;\Rightarrow$$

$$\hat{\mathbf{g}}_M(\mathbf{y}) = \arg\max_{\mathbf{g}} \underbrace{\mathcal{N}(\mathbf{y} - \Phi\mathbf{g}; 0, \lambda_{e_y} I)\, \mathcal{N}(\mathbf{g}, 0, K_g)}_{p(\mathbf{g},\mathbf{y}) = p(\mathbf{g}|\mathbf{y})p(\mathbf{y})}$$

$$\hat{\mathbf{g}}_A(\mathbf{y}) = \int \mathbf{g}\; p(\mathbf{g}|\mathbf{y})\; d\mathbf{g}$$

Example 1 cont'd

$$\mathbf{y}(\mathbf{g},\mathbf{e}_y) = \Phi\mathbf{g} + \mathbf{e}_y, \quad \mathbf{e}_y \sim \mathcal{N}(0, \lambda_{e_y} I),\ \mathbf{g} \sim \mathcal{N}(\bar{\mathbf{g}}, K_g)$$

$$\begin{bmatrix}\mathbf{g}\\ \mathbf{y}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\bar{\mathbf{g}}\\ \bar{\mathbf{g}}\end{bmatrix}, \begin{bmatrix}\Sigma_{gg} & \Sigma_{gy}\\ \Sigma_{yg} & \Sigma_{yy}\end{bmatrix}\right)$$

$$\text{where } \begin{bmatrix}\Sigma_{gg} & \Sigma_{gy}\\ \Sigma_{yg} & \Sigma_{yy}\end{bmatrix} = \begin{bmatrix}K_g & K_g\Phi^T\\ \Phi K_g & \Phi K_g \Phi^T + \lambda_{e_y} I\end{bmatrix}$$

From the theory of Gaussian rv:

$$p(\mathbf{g}|\mathbf{y}) = \mathcal{N}(\mathbf{g}; \mathbb{E}\{\mathbf{g}|\mathbf{y}\}, \mathrm{Cov}\{\mathbf{g}|\mathbf{y}\})$$

$$\mathbb{E}\{\mathbf{g}|\mathbf{y}\} = \Sigma_{gy}\Sigma_{yy}^{-1}(\mathbf{y} - \mathbb{E}\{\mathbf{y}\}) + \mathbb{E}\{\mathbf{g}\}$$

Both the *maximum of rankings estimate* and the *average ranking estimate* of $\mathbf{g}$ are thus given by

$$\hat{\mathbf{g}} = \Sigma_{gy}\Sigma_{yy}^{-1}(\mathbf{y} - \bar{\mathbf{g}}) + \bar{\mathbf{g}} = K_g\Phi^T\left(\Phi K_g \Phi^T + \lambda_{e_y} I\right)^{-1}(\mathbf{y} - \bar{\mathbf{g}}) + \bar{\mathbf{g}}$$

Special case: $\mathbf{y} = \mathbf{g} + \mathbf{e}_y$ $(\Phi = I)$, $K_g = \lambda_g I$

$$\hat{\mathbf{g}} = \frac{\lambda_g}{\lambda_g + \lambda_{e_y}}\mathbf{y} + \frac{\lambda_{e_y}}{\lambda_g + \lambda_{e_y}}\bar{\mathbf{g}} = \text{trust in data} \times \mathbf{y} + \text{trust in ranking} \times \bar{\mathbf{g}}$$

Estimating functions of unknowns

$$\boldsymbol{\theta} = f(\boldsymbol{\xi})$$

Estimates:

$$\hat{\boldsymbol{\theta}} = f(\hat{\boldsymbol{\xi}}_M), \quad \hat{\boldsymbol{\theta}} = f(\hat{\boldsymbol{\xi}}_A)$$

Alternatives:

$$\hat{\boldsymbol{\theta}}_M(\mathbf{z}) = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}; \mathbf{z})$$

$$p(\boldsymbol{\theta}; \mathbf{z}) := \int_{\boldsymbol{\Xi}(\mathbf{z}) \cap \{\boldsymbol{\xi} \in \boldsymbol{\Xi}:\, f(\boldsymbol{\xi}) = \theta\}} p(\boldsymbol{\xi}) d\boldsymbol{\xi}$$

Nuisance parameters have been marginalized (integrated) out

$$\hat{\boldsymbol{\theta}}_A(\mathbf{z}) := \frac{\int_{\boldsymbol{\Xi}(\mathbf{y})} f(\boldsymbol{\xi}) p(\boldsymbol{\xi})\, d\boldsymbol{\xi}}{p_y(\mathbf{y})} = \int f(\boldsymbol{\xi}) p(\boldsymbol{\xi}|\mathbf{z})\, d\boldsymbol{\xi} = \mathbb{E}\left\{ f(\boldsymbol{\xi}) | \mathbf{z} \right\}$$

Average over fs that are unfalsified

Outline

Choosing the ranking function $p(\boldsymbol{\xi})$

Notice that $\{\boldsymbol{\Xi}(\mathbf{z})\}_{\mathbf{z}}$ are disjoint sets ($M(\boldsymbol{\xi})$ single valued).

For given data $\mathbf{z}$, the ranking function is only used to rank the parameters in $\boldsymbol{\Xi}(\mathbf{z})$.

Thus we only need to choose the rankings for $\boldsymbol{\xi}$ in this set.

Common approach: Parameterized ranking $p = p(\boldsymbol{\xi}; \boldsymbol{\eta}(\mathbf{z}))$

How to determine the (hyper-) parameters $\boldsymbol{\eta}(\mathbf{z})$?

Let us use the rankings relevant for the data $\mathbf{z}$, $p(\boldsymbol{\xi}; \boldsymbol{\eta})$, $\boldsymbol{\xi} \in \boldsymbol{\Xi}(\mathbf{z})$, to compute rankings for $\boldsymbol{\eta}$:

i) Average ranking: $p_z(\mathbf{z}; \boldsymbol{\eta})$

ii) Optimistic ranking: $\sup_{\boldsymbol{\xi} \in \boldsymbol{\Xi}(\mathbf{z})} p(\boldsymbol{\xi}; \boldsymbol{\eta})$

How can we use the rankings of $\boldsymbol{\eta}$ for estimation of $\boldsymbol{\eta}$?

One possibility: $\boldsymbol{\eta}(\mathbf{z}) = \hat{\boldsymbol{\eta}}_{ML}(\mathbf{z}) := \arg\max_{\boldsymbol{\eta}} p_z(\mathbf{z}; \boldsymbol{\eta})$

Maximize the average of the rankings

Example 1 cont'd: Special case

$$\mathbf{y}(\mathbf{g}, \mathbf{e}_y) = \mathbf{g} + \mathbf{e}_y, \quad \mathbf{y} \in \mathbb{R}^N$$

$$\mathbf{e}_y \sim \mathcal{N}(0, \lambda_{e_y} I), \ \mathbf{g} \sim \mathcal{N}(\bar{\mathbf{g}}, \lambda_g I)$$

$$\begin{bmatrix} \mathbf{g} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \bar{\mathbf{g}} \\ \bar{\mathbf{g}} \end{bmatrix}, \begin{bmatrix} \lambda_g I & \lambda_g I \\ \lambda_g I & \lambda_g I + \lambda_{e_y} I \end{bmatrix} \right)$$

- λ_g does not directly influence the model $\mathbf{y}(\mathbf{g}, \mathbf{e}_y)$
- Such parameters are called *hyperparameters*
- The noise variance λ_{e_y} and $\bar{\mathbf{g}}$ are also hyperparameters but we will for simplicty assume them to be fixed.

$$-\log p(\mathbf{y}; \lambda_g) = \frac{1}{2}(\mathbf{y} - \bar{\mathbf{g}})^T \left(\lambda_g I + \lambda_{e_y} I\right)^{-1} (\mathbf{y} - \bar{\mathbf{g}}) + \frac{1}{2} \log\det\left(\lambda_g I + \lambda_{e_y} I\right)$$

$$= \frac{\|\mathbf{y} - \bar{\mathbf{g}}\|^2}{\lambda_g + \lambda_{e_y}} + N \log(\lambda_g + \lambda_{e_y})$$

Example 1 cont'd: Special case

$$\mathbf{y}(\mathbf{g}, \mathbf{e}_y) = \mathbf{g} + \mathbf{e}_y, \quad \mathbf{y} \in \mathbb{R}^N$$

$$\mathbf{e}_y \sim \mathcal{N}(0, \lambda_{e_y} I), \ \mathbf{g} \sim \mathcal{N}(\bar{\mathbf{g}}, \lambda_g I)$$

$$-\log p(\mathbf{y}; \lambda_g) = \frac{\|\mathbf{y} - \bar{\mathbf{g}}\|^2}{\lambda_g + \lambda_{e_y}} + N \log(\lambda_g + \lambda_{e_y})$$

Estimate

$$\hat{\lambda}_g = \frac{1}{N} \|\mathbf{y} - \bar{\mathbf{g}}\|^2 - \lambda_{e_y}$$

Spread of $\mathbf{y}$ around $\bar{\mathbf{g}}$, accounting for spread of $\mathbf{e}_y$.

$$\hat{\mathbf{g}}(\hat{\lambda}_g) = \frac{\hat{\lambda}_g}{\hat{\lambda}_g + \lambda_{e_y}} \mathbf{y} = \left(1 - \frac{\lambda_{e_y}}{\frac{1}{N}\|\mathbf{y} - \bar{\mathbf{g}}\|^2}\right) \mathbf{y} + \frac{\lambda_{e_y}}{\frac{1}{N}\|\mathbf{y} - \bar{\mathbf{g}}\|^2} \bar{\mathbf{g}}$$

Example 1 cont'd: Special case

$$\mathbf{y}(\mathbf{g}, \mathbf{e}_y) = \mathbf{g} + \mathbf{e}_y, \quad \mathbf{y} \in \mathbb{R}^N$$

$$\mathbf{e}_y \sim \mathcal{N}(0, \lambda_{e_y} I), \ \mathbf{g} \sim \mathcal{N}(\bar{\mathbf{g}}, \lambda_g I)$$

ML-estimate

$$\hat{\lambda}_g = \frac{1}{N}\|\mathbf{y} - \bar{\mathbf{g}}\|^2 - \lambda_{e_y}$$

$$\hat{\mathbf{g}}(\hat{\lambda}_g) = \left(1 - \frac{\lambda_{e_y}}{\frac{1}{N}\|\mathbf{y} - \bar{\mathbf{g}}\|^2}\right)\mathbf{y} + \frac{\lambda_{e_y}}{\frac{1}{N}\|\mathbf{y} - \bar{\mathbf{g}}\|^2}\bar{\mathbf{g}}$$

Interpretation:

- With $\mathbf{g}$ fix, $\mathbf{y} \sim \mathcal{N}(\mathbf{g}, \lambda_{e_y} I)$
- Hypothesis H_o: $\mathbf{g} = \bar{\mathbf{g}}$
- Under H_o, $T := \|\mathbf{y} - \bar{g}\|^2/\lambda_{e_y} \sim \chi^2(N)$
- Under H_o: $\mathbb{E}\{T\} = N \Rightarrow \hat{\mathbf{g}}(\hat{\lambda}_g) \approx \bar{g}$ if H_o true
- Hypothesis violated (T large) $\Rightarrow$ Data used

Exercise

- λ_{e_y} estimated as well $\Rightarrow$ James-Stein estimator
- James-Stein estimator outperforms ML
- As does our estimator

Let for simplicity $\bar{\mathbf{g}} = 0$ so that

$$\hat{\mathbf{g}}(\hat{\lambda}_g) = \left(1 - \frac{\lambda_{e_y}}{\frac{1}{N}\|\mathbf{y}\|^2}\right)\mathbf{y}$$

Take 5 min and think if it makes sense that this estimator beats the ML estimator

$$\hat{g}_{ML} = \mathbf{y}$$

in terms of the MSE
Starting point: $\mathbf{y} \sim \mathcal{N}(\mathbf{g}, \lambda_e I)$

Example 1 cont'd

$$\mathbf{y}(\mathbf{g}, \mathbf{e}_y) = \Phi \mathbf{g} + \mathbf{e}_y$$

Let instead

$$p(\mathbf{g}, \mathbf{e}_y) = \mathcal{N}(\mathbf{e}_y; 0, \lambda_{e_y} I)\delta(\mathbf{g} - \boldsymbol{\eta})$$

$\Rightarrow$ $\mathbf{g}$ is a singleton $\boldsymbol{\eta}$ which is to be determined from data.

$$\boldsymbol{\Xi}(\mathbf{y}) = \{(\mathbf{e}_y, \mathbf{g}) : \ \mathbf{y}(\mathbf{g}, \mathbf{e}_y) = \mathbf{y}\} = (\mathbf{y} - \Phi\boldsymbol{\eta}, \boldsymbol{\eta}) \text{ singleton}$$

$$p_y(\mathbf{y}; \mathbf{g}) := \mathcal{N}(\mathbf{y} - \Phi \mathbf{g}; 0, \lambda_{e_y} I)$$

- $\hat{\mathbf{g}}_M(\mathbf{y}) = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}$
- In our special case $\Phi = I$, $\hat{\mathbf{g}}_M(\mathbf{y}) = \mathbf{y}$

Outline

Summary

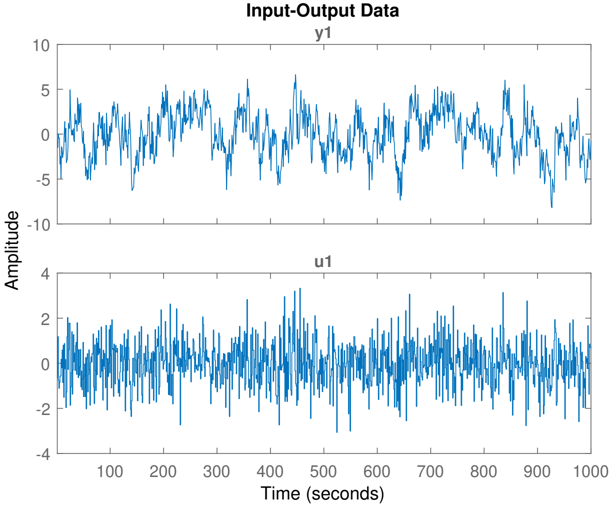
- Constructive model $\mathbf{z}(\boldsymbol{\xi})$, parametrized by vector of unknowns $\boldsymbol{\xi} \in \boldsymbol{\Xi}$
- Among the set of unfalsified parameters, the ranking determines the estimate
- Different functions can be used for this, e.g. average and maximum.
- Ranking function can also be parametrized $(\boldsymbol{\eta})$
- $\boldsymbol{\eta}$ can be estimated using the ranking function as well
  - Elements of $\boldsymbol{\eta}$ directly mapped to elements of ξ are usually referred to as model parameters, cf. $\mathbf{g}$ in Example 1.
  - Elements of $\boldsymbol{\eta}$ not directly mapped to elements of ξ are usually referred to as hyper-parameters, cf. λ_g in Example 1.
- Computations requires integration and optimization

Model simulation

Model

$$y(t) = \frac{bq^{-1}}{1 + fq^{-1}} u(t)$$

Data



Model simulation

b and f determined by minimizing

$$\sum_{t=1}^{N} (y(t) - \hat{y}(t, b, f))^2$$

$$\hat{y}(t; b, f) := \frac{bq^{-1}}{1 + fq^{-1}} u(t)$$

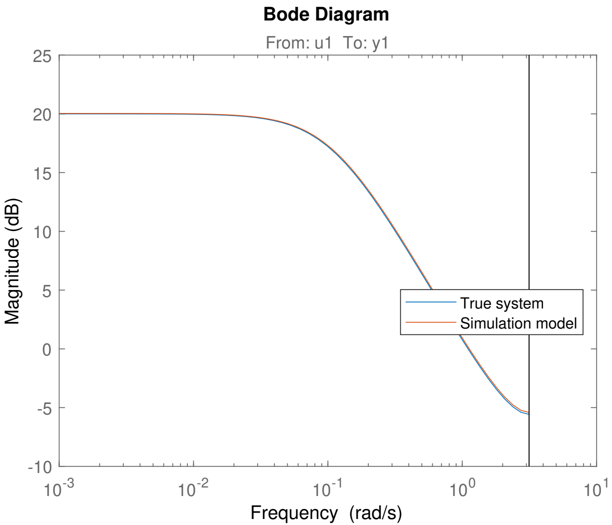
Computed from

$$(1 + fq^{-1})\hat{y}(t; b, f) = bq^{-1}u(t)$$

that is

$$\begin{aligned}
\hat{y}(t; b, f) &= -f\hat{y}(t-1, b, f) + bu(t-1) \\
\hat{y}(1; b, f) &= 0 \\
\vdots \quad & \quad \vdots \\
\hat{y}(5; b, f) &= -f^3bu(1) + f^2bu(2) - fbu(3) + bu(4) \\
\vdots \quad & \quad \vdots
\end{aligned}$$

Model simulation



Model simulation

$$(1 + fq^{-1})\hat{y}(t) = bq^{-1}u(t)$$

Very nonlinear optimization problem. Can we simplify?
Our model

$$(1 + fq^{-1})y(t) = bq^{-1}u(t)$$

can be written as

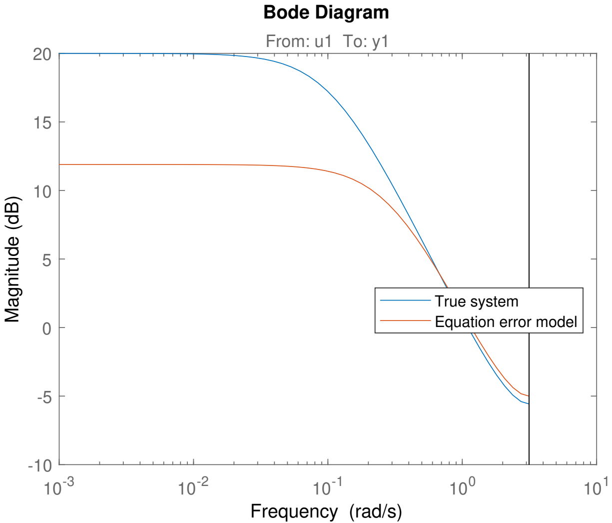
$$y(t) = -fy(t-1) + bu(t-1)$$

Take $\hat{y}(t) = -fy(t-1) + bu(t-1) \Rightarrow$ Minimize

$$\sum_{t=1}^{N}(y(t) - fy(t-1) - bu(t-1))^2$$

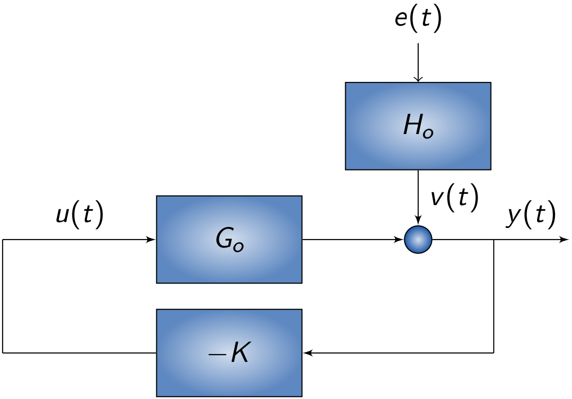
Least-squares problem!!!

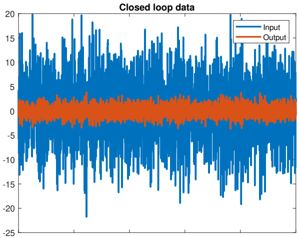
Model simulation



Why different results. Which one to use?

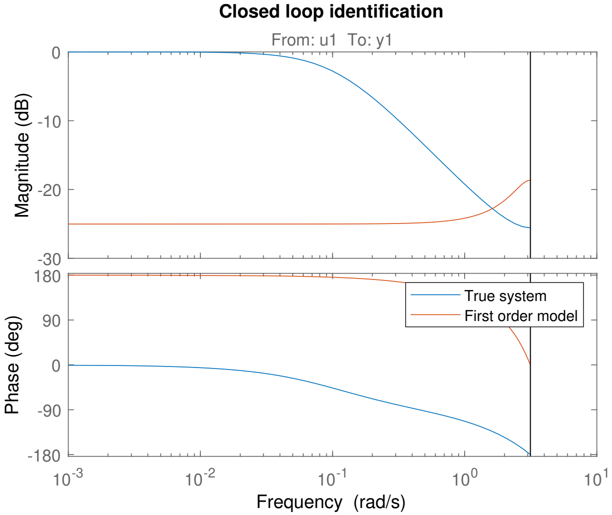
Closed loop identification





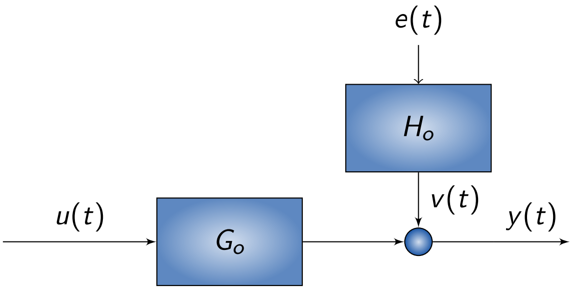
Closed loop identification

Result:

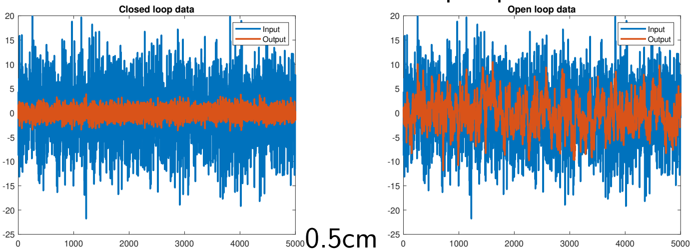


Closed loop identification

Open loop identification



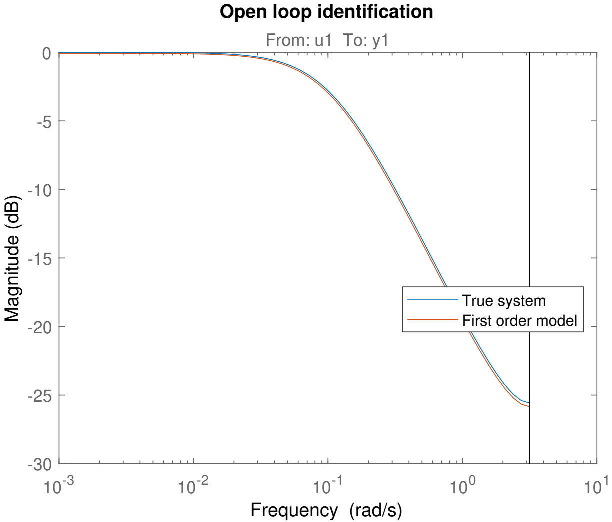
Data same characteristics as in closed loop experiment:



0.5cm

Closed loop identification

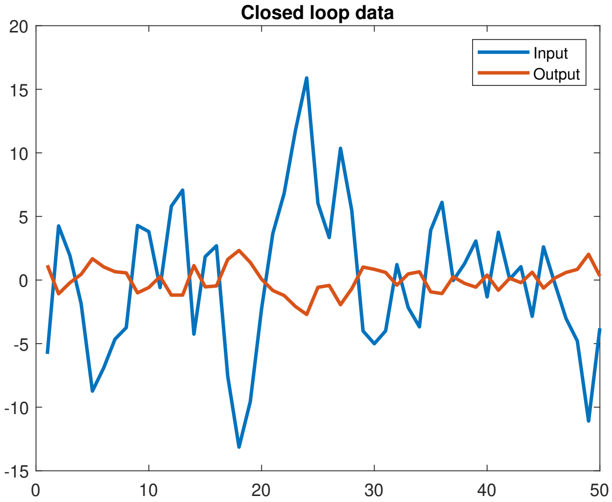
Result



What so peculiar about closed loop identification?

Closed loop identification

Close up

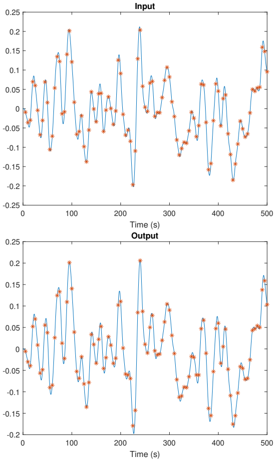


Opposite response to the eye!

Sampling

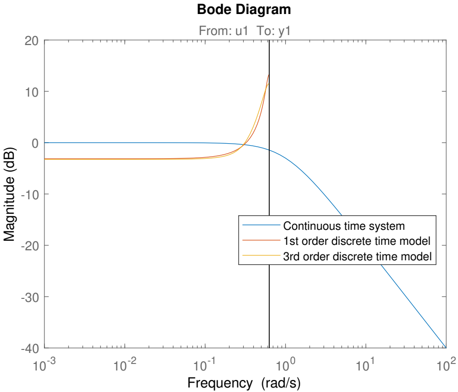
$$G(s) = \frac{1}{s+1}$$

Data:



Sampling

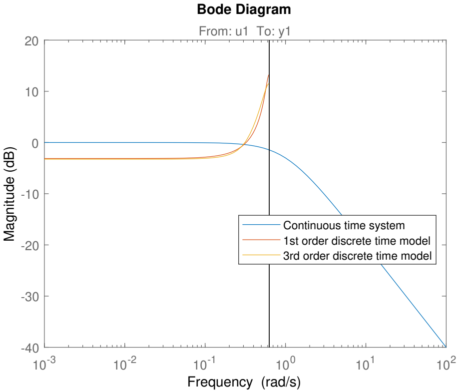
$$y(nT) = \frac{\sum_{k=1}^{n} b_k q^{-k}}{1 + \sum_{k=1}^{n} f_k q^{-k}} u(nT)$$



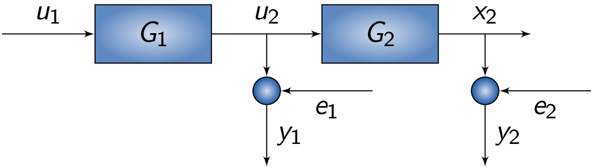
Noise free data, fast sampling. Yet problem???

Sampling

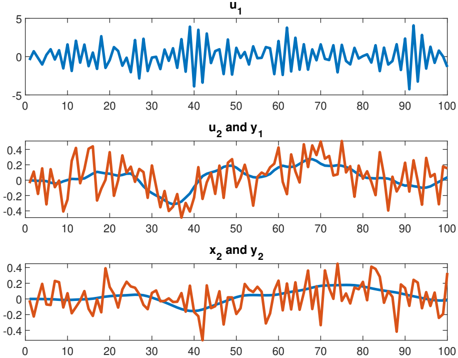
$$y(nT) = \frac{\sum_{k=1}^{n} b_k q^{-k}}{1 + \sum_{k=1}^{n} f_k q^{-k}} u(nT)$$



Measurement errors

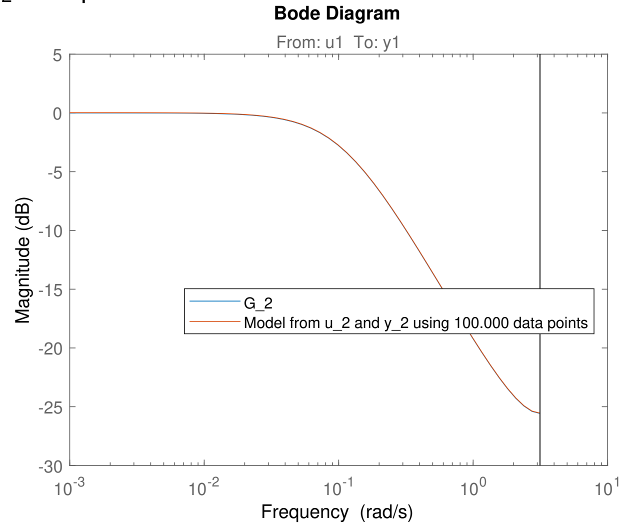


Interested in G_2 but also G_1 (high order) unknown
Large data set (100.000 samples). First 1000 shown



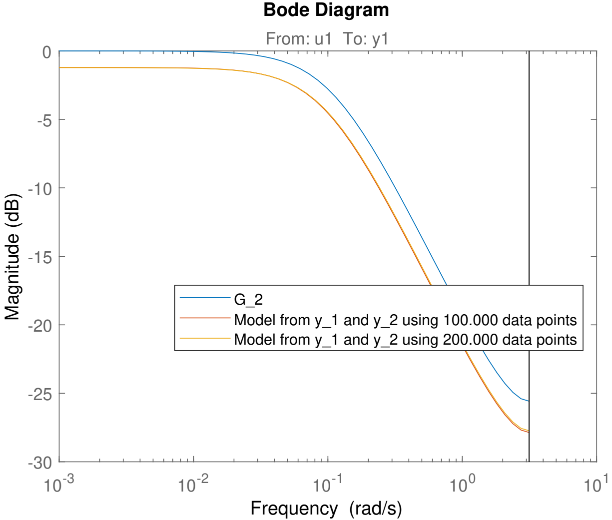
Measurement errors

Using u_2 as input



Measurement errors

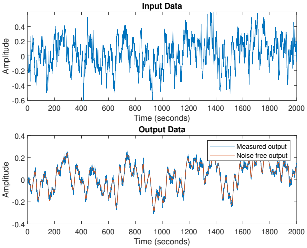
Using y_1 as input in the model



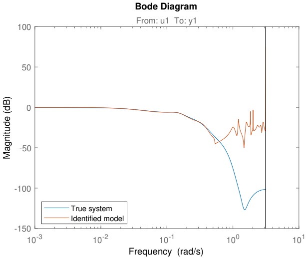
How handle measurement errors on inputs?

Complex models

System of known order 25

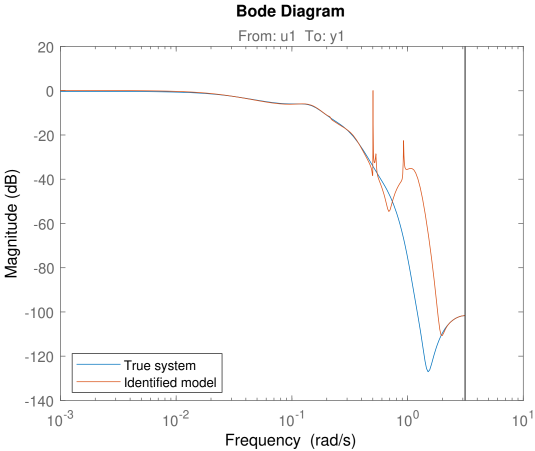


State-of-the art:



Complex models

Recall: Highly non-linear optimization problem. Need good inital values. Let us start at true values.



Still problems. How to deal with complex systems?

Hilbert spaces

Let $\mathcal{V}$ be a vector space equipped with an inner product $\langle \cdot, \cdot \rangle$

1. $\langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle$
2. $\langle \lambda u, v \rangle = \lambda \langle u, v \rangle$
3. $\langle u, v \rangle = \langle v, u \rangle^*$
4. $\langle v, v \rangle \geq 0$ with equality iff $v = 0$

Norm: $\|v\| = \sqrt{\langle v, v \rangle}$

Hilbert space $\mathcal{H}$: Complete inner product space (Cauchy sequences converge)

Extend definition to column vectors u and v of elements of $\mathcal{H}$:

$$\langle u, v \rangle = M, \quad M_{i,j} = \langle u_i, v_j \rangle$$

Example 1: Consider the columns of $X \in \mathbb{R}^{N \times n_x}$ and $Y \in \mathbb{R}^{N \times n_y}$ as elements of $\mathbb{R}^N$, then

$$\langle X, Y \rangle = X^T Y$$

Example 2: Let $\mathbf{x} \in \mathbb{R}^{n_x}$ and $\mathbf{y} \in \mathbb{R}^{n_y}$ be random vectors with finite second moments. Then

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbb{E}\left\{\mathbf{x}\mathbf{y}^T\right\}$$

Orthogonal projections

Orthogonality

An element $u \in \mathcal{H}$ is orthogonal to the subspace $\mathcal{S} \subseteq \mathcal{H}$ if

$$\langle u, v \rangle = 0 \quad \forall v \in \mathcal{S}.$$

We write $u \perp \mathcal{S}$

Projection theorem

Let $u \in \mathcal{H}$ be given and let $\mathcal{S} \subseteq \mathcal{H}$ be a closed subspace to $\mathcal{H}$. Then there exists a unique $v \in \mathcal{S}$ such that $u - v \perp \mathcal{S}$. The element v is the unique solution to

$$\min_{v \in \mathcal{S}} \|u - v\|$$

v is called the orthogonal projection of u onto $\mathcal{S}$ and is denoted $u_{\mathcal{S}}$

It follows that $u \in \mathcal{H}$ has a unique decomposition

$u = u_{\mathcal{S}} + u_{\mathcal{S}^\perp}$, where $u_{\mathcal{S}^\perp} = u - u_{\mathcal{S}} \in \mathcal{S}^\perp$ (subspace orthogonal to $\mathcal{S}$)

Orthogonal projections: Pythagoras relation

$$u = u_{\mathcal{S}} + u_{\mathcal{S}^\perp} \;\Rightarrow\; \|u\|^2 = \|u_{\mathcal{S}}\|^2 + \|u_{\mathcal{S}^\perp}\|^2$$

In our context often written as

$$\|u\|^2 - \|u_{\mathcal{S}}\|^2 = \|u_{\mathcal{S}^\perp}\|^2 = \|u - u_{\mathcal{S}}\|^2$$

The projection theorem:

$$\|u - v\|^2 \geq \|u - u_{\mathcal{S}}\|^2 = \|u_{\mathcal{S}^\perp}\|^2 = \|u\|^2 - \|u_{\mathcal{S}}\|^2 \geq 0 \quad \forall v \in \mathcal{S}$$

Vector version:

$$\langle u - v, u - v\rangle \geq \langle u - u_{\mathcal{S}}, u - u_{\mathcal{S}}\rangle = \langle u, u\rangle - \langle u_{\mathcal{S}}, u_{\mathcal{S}}\rangle \geq 0 \quad \forall v \in \mathcal{S}$$

Matrix inequality

Note: Projection $u_{\mathcal{S}}$ has smaller "norm" than u: $\langle u, u\rangle - \langle u_{\mathcal{S}}, u_{\mathcal{S}}\rangle \geq 0$

Orthogonal projections: Finite dimensional subspaces

Problem: Project all elements of the n_u-dimensional vector u on the linear span of the elements of the vector y (solve n_u projections simultaneously)

$\mathcal{S} = \{Ly : \; L \in \mathbb{R}^{n_u \times n_y}\}$

Optimality condition:

$$\begin{aligned} 0 &= \langle u - Ly, \mathbf{y} \rangle = \langle u, y \rangle - L\langle y, y \rangle \\ &\Rightarrow \; L^* = \langle u, y \rangle \langle y, y \rangle^{-1} \\ &\Rightarrow \; u_{\mathcal{S}} = L^* y = \langle u, y \rangle \langle y, y \rangle^{-1} y \end{aligned}$$

Projection theorem and Pythagoras: $v = Ly \Rightarrow$

$$\langle u - v, u - v \rangle \geq \langle u - L^* y, u - L^* y \rangle = \langle u, u \rangle - \langle u, y \rangle \langle y, y \rangle^{-1} \langle y, u \rangle$$

Example: Rows of $U \in \mathbb{R}^{n_u \times N}$ to be projected on the rows of $Y \in \mathbb{R}^{n_y \times N}$

$$\begin{aligned} U_{\mathcal{S}} &= U^T Y (Y^T Y)^{-1} Y \\ 0 &\geq (U - U_{\mathcal{S}})^T (U - U_{\mathcal{S}}) = U^T U - U^T Y (Y^T Y)^{-1} Y^T U \end{aligned}$$

Summary