

Data Driven Modeling

Lecture 1

Håkan Hjalmarsson

KTH - ROYAL INSTITUTE OF TECHNOLOGY

hjalmars@kth.se

May 15, 2020



Outline

Introduction

- Practicalities

- Outline

Signals

- Continuous time signals

- Discrete time signals

Dynamic systems

Introduction to parameter estimation

- Some examples

- Key problem

- Choosing the ranking function

- Summary

Inspiring pitfalls

Outline

Introduction

- Practicalities

- Outline

Signals

- Continuous time signals

- Discrete time signals

Dynamic systems

Introduction to parameter estimation

- Some examples

- Key problem

- Choosing the ranking function

- Summary

Inspiring pitfalls

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. Preferably 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

Introduction

Practicalities

Outline

Signals

Continuous time signals

Discrete time signals

Dynamic systems

Introduction to parameter estimation

Some examples

Key problem

Choosing the ranking function

Summary

Inspiring pitfalls

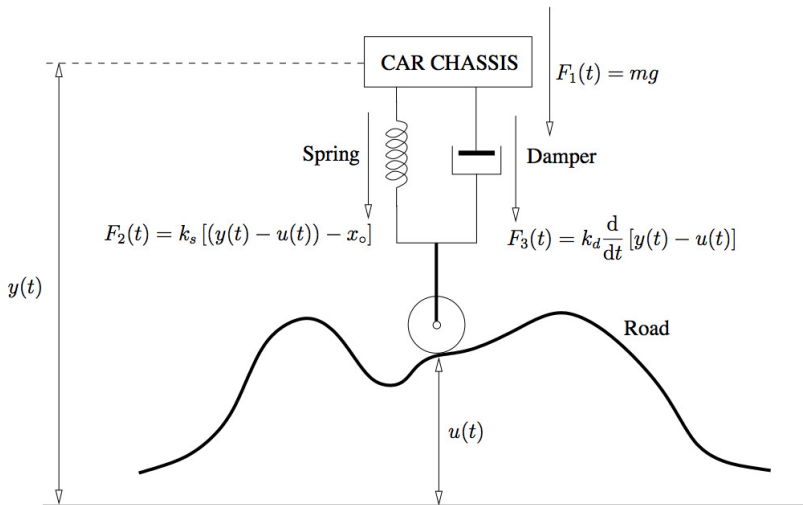
Course Outline

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

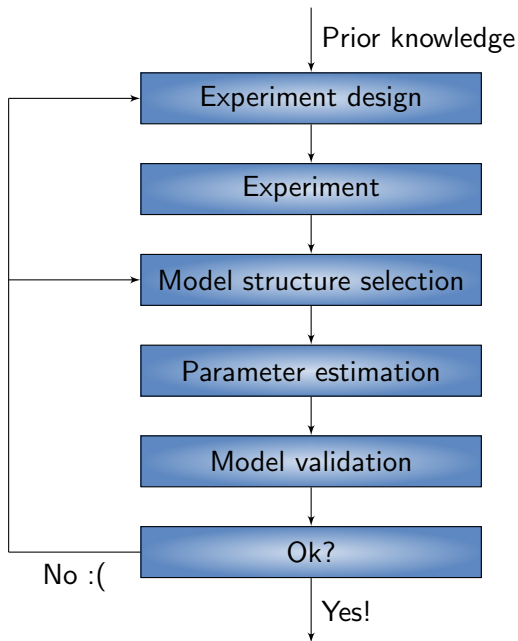
Course Outline

5. Biased parameter estimation (Tuesday 2/6, 10-12) . Chapter 6 in LN.
 - o The bias-variance trade-off
 - o The Cramér-Rao lower bound
 - o Average risk minimization
 - o Minimax estimation
 - o Pointwise risk minimization
6. Asymptotic theory (Friday 5/6, 15-17). Chapter 7 in L.N. Chapter 8 in Ljung
 - o Limits of random variables
 - o Large sample properties of estimators
 - o Using estimation for parameter estimation, part II
 - o Large sample properties of biased estimators
7. Computational aspects (Tuesday 9/6, 08-10). Chapter 10 in Ljung.
 - o Gradient based optimization
 - o Convex relaxations
 - o Integration by Markov Chain Monte Carlo (MCMC) methods
8. Case studies I (Friday 12/6, 10-12)
 - o Parametric LTI models
 - o Impulse response models
9. Case studies II (Tuesday 16/6, 10-12)
 - o Uncertain input models
 - o 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

Introduction

Practicalities

Outline

Signals

Continuous time signals

Discrete time signals

Dynamic systems

Introduction to parameter estimation

Some examples

Key problem

Choosing the ranking function

Summary

Inspiring pitfalls

Continuous time signals

Definition

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

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

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

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

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

Continuous time signals

Fourier transform and its inverse

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

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 .

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

Outline

Introduction

Practicalities

Outline

Signals

Continuous time signals

Discrete time signals

Dynamic systems

Introduction to parameter estimation

Some examples

Key problem

Choosing the ranking function

Summary

Inspiring pitfalls

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)

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

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

Discrete time signals

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

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

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

Theorem

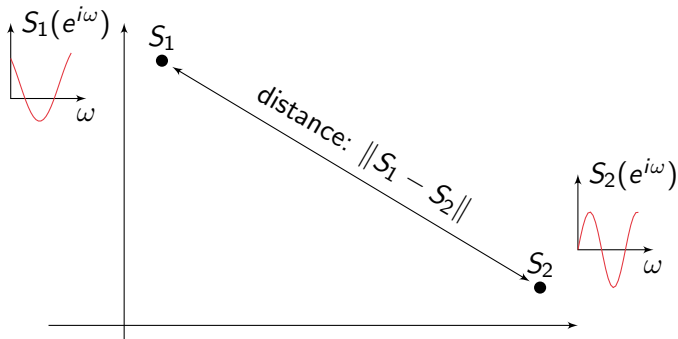
Any $S \in L_2(\mathbb{T})$ can be represented as $S(e^{i\omega}) = \sum_{t=-\infty}^{\infty} s(t) e^{-i\omega t}$ where

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

What does $S = 0$ mean in $L_2(\mathbb{T})$?

Discrete time signals

What does $S = 0$ mean in $L_2(\mathbb{T})$?



Functions for which

$$0 = \|S_1 - S_2\|_2^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |S_1(e^{i\omega}) - S_2(e^{i\omega})|^2 d\omega$$

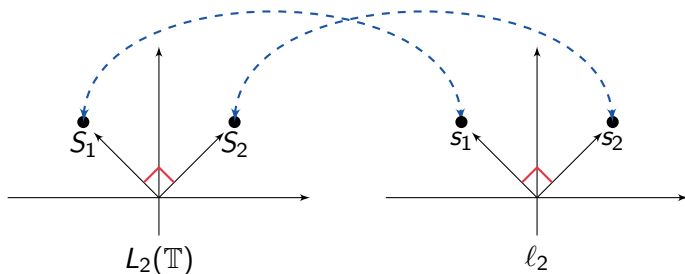
mapped to same point in $L_2(\mathbb{T})$

Called equivalence classes.

Discrete time signals

ℓ_2 and $L_2(\mathbb{T})$ isomorphic: 1-1 relationship between elements

Isomorphism



Geometric properties preserved! \Leftrightarrow Parseval's formula

$$\langle S, V \rangle = \langle s, v \rangle$$

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

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

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

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

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

For $S = \sum_{t=0}^{\infty} \bar{s}(t)z^{-t}$ holomorphic in $|z| > 1$:

$$S \in H_2 \Leftrightarrow \sum |\bar{s}(t)|^2 < \infty$$

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. $\ell_\infty \rightarrow \ell_\infty$. 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

$$x(t+1) = A(\theta)x(t) + B(\theta)u(t) + K(\theta)e(t)$$

$$y(t) = C(\theta)x(t) + D(\theta)u(t) + e(t)$$

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

- Non-linear

$$x(t+1) = f(x(t), u(t), w(t), \theta)$$

$$y(t) = h(x(t), u(t), e(t), \theta)$$

Common linear black-box structures

- FIR

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

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

- FIR

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

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.

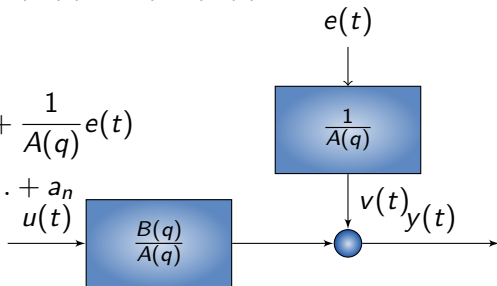
Common linear black-box structures

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

- ARX

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

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



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

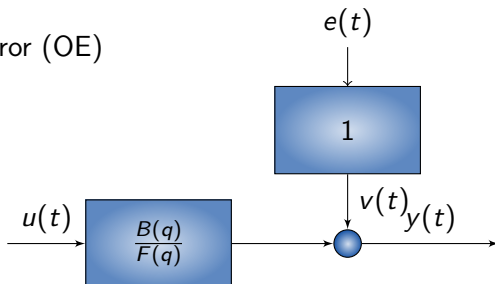
$$y(t) = \varphi^T \theta + e(t)$$

$$\varphi(t) = [-y(t-1) \quad \dots \quad -y(t-n) \quad u(t-1) \quad \dots \quad u(t-n)]^T$$

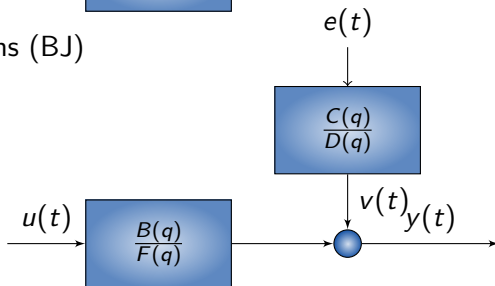
$$\theta = [a_1 \quad \dots \quad a_n \quad b_1 \quad \dots \quad b_n]^T$$

Common linear black-box structures

- Output-Error (OE)



- Box-Jenkins (BJ)



Continuous time models

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

Sampling gives

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

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.

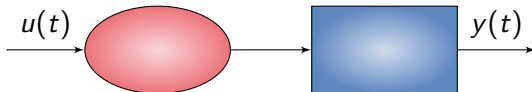
- ▶ Neural networks
- ▶ Radial basis functions
- ▶ NLARX: $\varphi(t)$ past inputs and outputs
- ▶
- ▶
- ▶

- Block oriented models

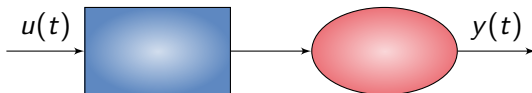
Block-oriented models



- Hammerstein (nonlinear actuator)



- Wiener (nonlinear sensor)



- Hammerstein-Wiener



Outline

Introduction

Practicalities

Outline

Signals

Continuous time signals

Discrete time signals

Dynamic systems

Introduction to parameter estimation

Some examples

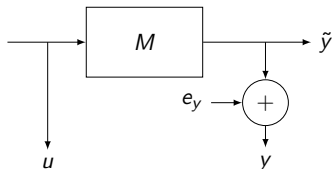
Key problem

Choosing the ranking function

Summary

Inspiring pitfalls

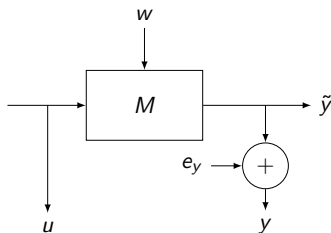
Example 1: Scalar LTI model



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

- 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

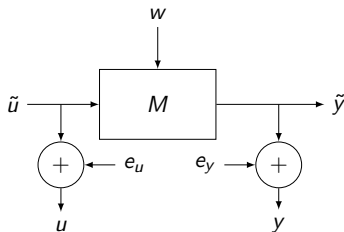


$$\mathbf{x} = F(\boldsymbol{\theta})\mathbf{u} + G(\boldsymbol{\theta})\mathbf{w}$$

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

- 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



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

$$\mathbf{u} = \tilde{\mathbf{u}} + \mathbf{e}_u$$

$$\mathbf{y} = H(\theta)\mathbf{x} + \mathbf{e}_y$$

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

Outline

Introduction

Practicalities

Outline

Signals

Continuous time signals

Discrete time signals

Dynamic systems

Introduction to parameter estimation

Some examples

Key problem

Choosing the ranking function

Summary

Inspiring pitfalls

Key issue #1: More unknowns than measurements

Collect all unknowns in $\xi \in \Xi$.

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

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

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

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

Maximum of rankings estimate:

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

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:

$$\delta(\mathbf{x}) := \prod_{k=1}^n \delta(x(k)), \quad \mathbf{x} = [x(1) \quad \dots \quad x(n)]^T \in \mathbb{R}^n$$

The joint ranking of model parameters ξ and observations \mathbf{z} :

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

Gives:

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

Key issue #1: More unknowns than measurements

Alternative: *Average of rankings estimate*:

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

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

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

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

Side effect:

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

Both estimates obtained from $p(\xi|\mathbf{z})$

That's it folks - the course is finished!

From here on it can only become more confusing

Example 1 cont'd

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

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

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

Maximum ranking model:

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

$$\begin{aligned} & \underbrace{\mathcal{N}(\mathbf{e}_y; 0, \lambda_{e_y} I) \mathcal{N}(\mathbf{g}, 0, K_g)}_{p(\mathbf{g}, \mathbf{e}_y)} \\ &= \arg \max_{\xi} \underbrace{\mathcal{N}(\mathbf{e}_y; 0, \lambda_{e_y} I) \mathcal{N}(\mathbf{g}, 0, K_g) \delta(\mathbf{y} - \Phi \mathbf{g} - \mathbf{e}_y)}_{p(\mathbf{g}, \mathbf{e}_y, \mathbf{y})} \end{aligned}$$

Eliminating \mathbf{e}_y : $\mathbf{e}_y = \mathbf{y} - \Phi \mathbf{g} \Rightarrow$

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

Example 1 cont'd

Special case: $\Phi = I$, $K_g = \lambda_g I$, i.e.

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

skipping math:

$$\hat{\mathbf{g}}_M = \frac{\lambda_g}{\lambda_g + \lambda_{e_y}} \mathbf{y} + \frac{\lambda_{e_y}}{\lambda_g + \lambda_{e_y}} \bar{\mathbf{g}}$$

- $\lambda_{e_y} \gg \lambda_g$: A wide set of \mathbf{e}_y have about the same ranking, while a narrower set of \mathbf{g} have same ranking.

Trust initial ranking of \mathbf{g} , i.e. trust the highest ranked \mathbf{g} a priori, i.e. $\bar{\mathbf{g}}$

- $\lambda_{e_y} \ll \lambda_g$: A narrow set of \mathbf{e}_y have about the same ranking, while a larger set of \mathbf{g} have same ranking.

Trust data \mathbf{y}

Example 1 cont'd

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

Average of rankings estimate:

$$\begin{aligned}\hat{\mathbf{g}}_A(\mathbf{y}) &= \frac{\int \int \mathbf{g} p(\mathbf{g}, \mathbf{e}_y | \mathbf{y}) d\mathbf{e}_y d\mathbf{g}}{p(\mathbf{y})} \\&= \frac{\overbrace{\int \mathbf{g} \int \mathcal{N}(\mathbf{e}_y; 0, \lambda_{e_y} I) \mathcal{N}(\mathbf{g}, 0, K_g) \delta(\mathbf{y} - \Phi \mathbf{g} - \mathbf{e}_y) d\mathbf{e}_y d\mathbf{g}}^{\text{Total ranking for the pair } (\mathbf{g}, \mathbf{y}), \text{ i.e. } p(\mathbf{g}, \mathbf{y})}}{p(\mathbf{y})} \\&= \frac{\int \mathbf{g} \overbrace{\mathcal{N}(\mathbf{y} - \Phi \mathbf{g}; 0, \lambda_{e_y} I) \mathcal{N}(\mathbf{g}, 0, K_g)}^{p(\mathbf{g}, \mathbf{y})} d\mathbf{g}}{p(\mathbf{y})} \\&= \int \mathbf{g} p(\mathbf{g} | \mathbf{y}) d\mathbf{g}, \text{ where } p(\mathbf{g} | \mathbf{y}) = p(\mathbf{g}, \mathbf{y}) / p(\mathbf{y})\end{aligned}$$

Special case (skipping math): $\Phi = I$, $K_g = \lambda_g I$. Same as $\hat{\mathbf{g}}_M$.

Estimating functions of unknowns

$$\theta = f(\xi)$$

Estimates:

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

Alternatives:

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

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

Nuisance parameters have been marginalized (integrated) out

$$\hat{\theta}_A(\mathbf{z}) := \frac{\int_{\Xi(\mathbf{y})} f(\xi) p(\xi) d\xi}{p_{\mathbf{y}}(\mathbf{y})} = \int f(\xi) p(\xi | \mathbf{z}) d\xi = \mathbb{E}[f(\xi) | \mathbf{z}]$$

Average over f s that are unfalsified

Outline

Introduction

Practicalities

Outline

Signals

Continuous time signals

Discrete time signals

Dynamic systems

Introduction to parameter estimation

Some examples

Key problem

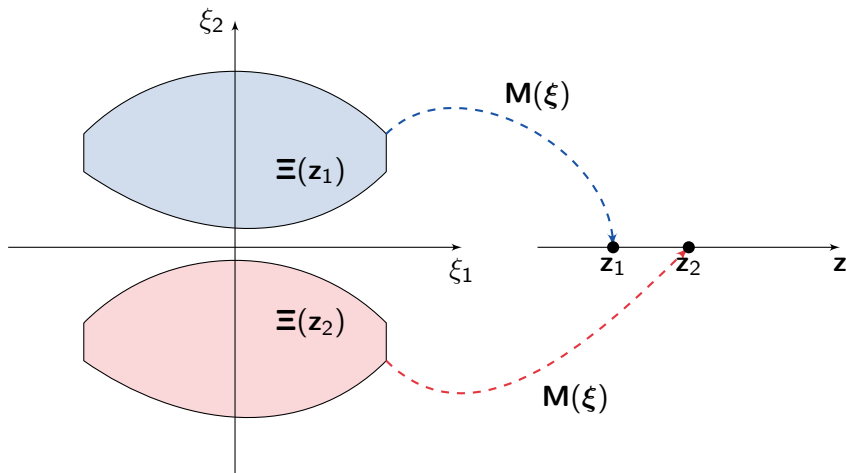
Choosing the ranking function

Summary

Inspiring pitfalls

Choosing the ranking function

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



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

Choosing the ranking function

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

Thus we only need to choose the rankings for ξ in this set.

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

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

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

i) Total ranking: $p_z(\mathbf{z}; \eta) = \int p(\xi, \mathbf{z}; \eta) d\xi$

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

How can we use the rankings of η for estimation of η ?

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

Maximize the total of the rankings

Example 1 cont'd: Special case

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

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

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

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

$$\begin{aligned} -\log p(\mathbf{y}; \lambda_g) &= \frac{1}{2} (\mathbf{y} - \bar{\mathbf{g}})^T (\lambda_g I + \lambda_{e_y} I)^{-1} (\mathbf{y} - \bar{\mathbf{g}}) + \frac{1}{2} \log \det (\lambda_g I + \lambda_{e_y} I) \\ &= \frac{\|\mathbf{y} - \bar{\mathbf{g}}\|^2}{\lambda_g + \lambda_{e_y}} + N \log(\lambda_g + \lambda_{e_y}) \end{aligned}$$

Example 1 cont'd: Special case

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

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

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

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 .

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

Example 1 cont'd: Special case

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

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

ML-estimate

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

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

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

Example 1 cont'd

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

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.

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

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

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

Outline

Introduction

Practicalities

Outline

Signals

Continuous time signals

Discrete time signals

Dynamic systems

Introduction to parameter estimation

Some examples

Key problem

Choosing the ranking function

Summary

Inspiring pitfalls

Summary

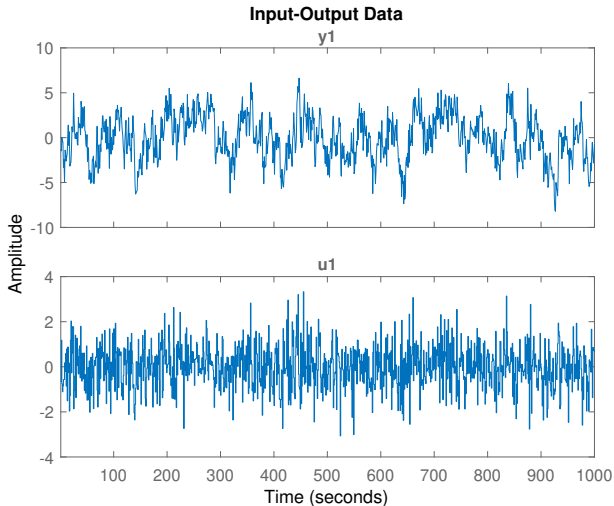
- Constructive model $\mathbf{z}(\xi)$, parametrized by vector of unknowns $\xi \in \Xi$
- Among the set of unfalsified parameters, the ranking determines the estimate
- Different functions can be used for this, e.g. average and maximum.
- Ranking function can also be parametrized (η)
- η can be estimated using the ranking function as well
 - ▶ Elements of η directly mapped to elements of ξ are usually referred to as model parameters, cf. \mathbf{g} in Example 1.
 - ▶ Elements of η 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

$$\hat{y}(t; b, f) = -f\hat{y}(t-1, b, f) + bu(t-1)$$

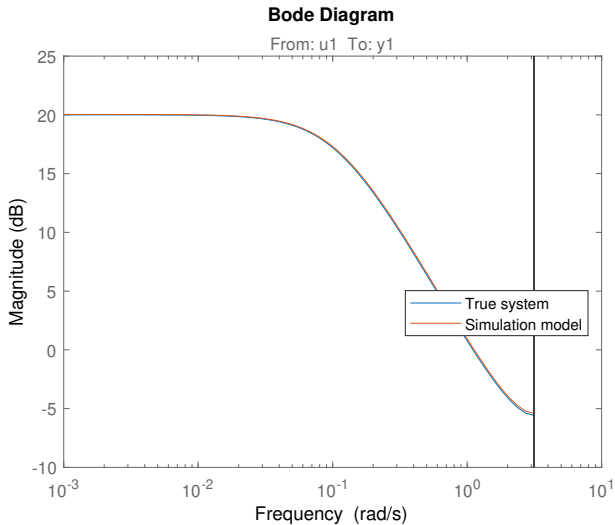
$$\hat{y}(1; b, f) = 0$$

$$\vdots \quad \quad \vdots$$

$$\hat{y}(5; b, f) = -f^3 bu(1) + f^2 bu(2) - f bu(3) + bu(4)$$

$$\vdots \quad \quad \vdots$$

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

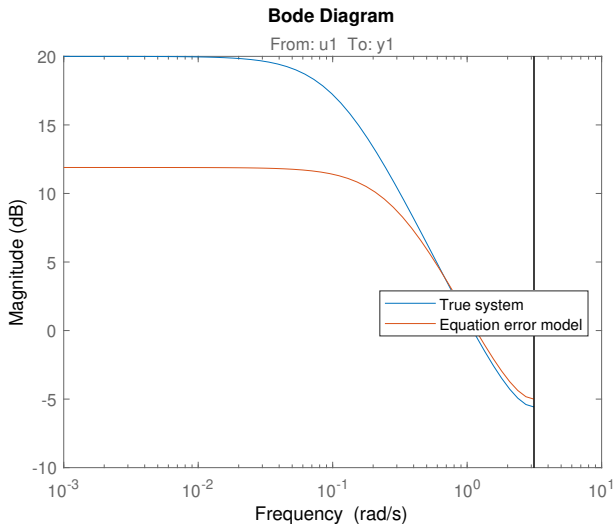
$$y(t) = -fy(t-1) + bu(t-1)$$

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

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

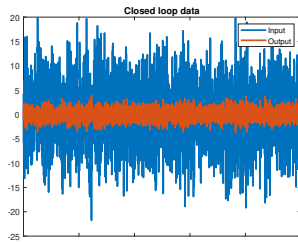
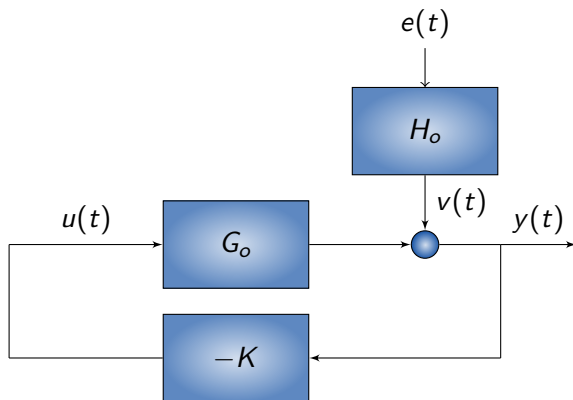
Least-squares problem!!!

Model simulation



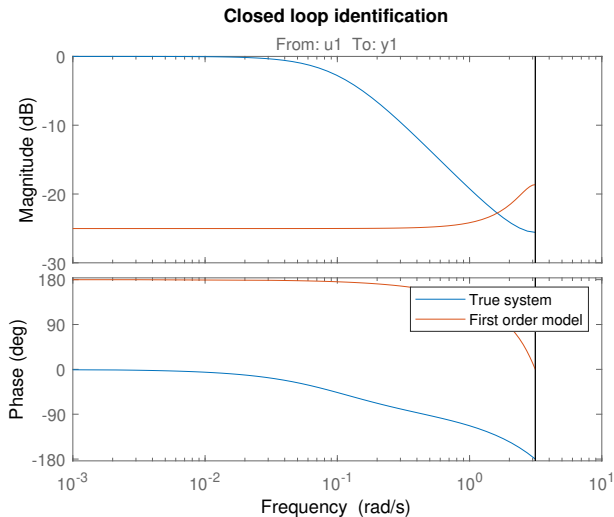
Why different results. Which one to use?

Closed loop identification



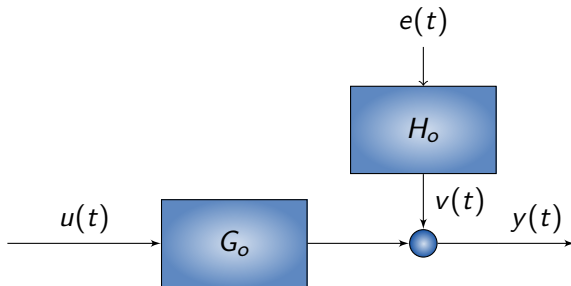
Closed loop identification

Result:

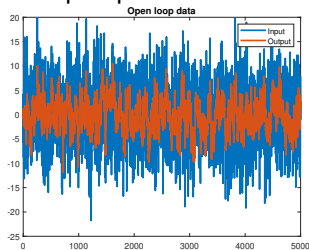
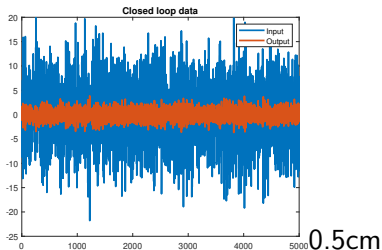


Closed loop identification

Open loop identification



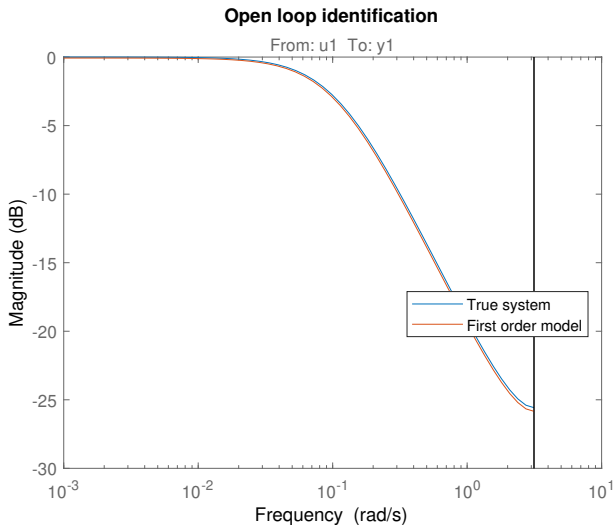
Data same characteristics as in closed loop experiment:



0.5cm

Closed loop identification

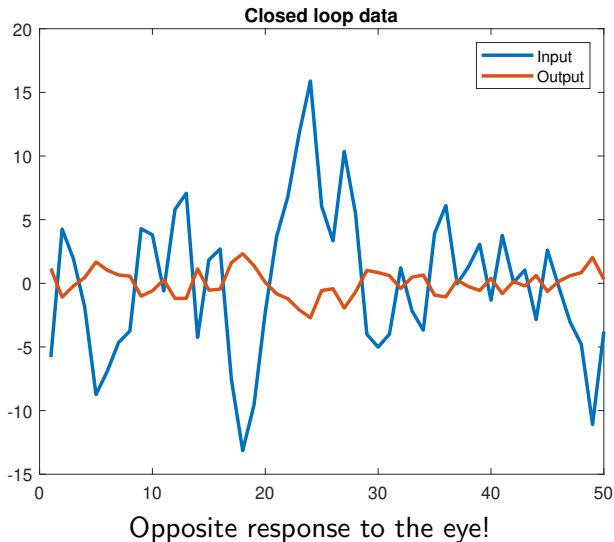
Result



What so peculiar about closed loop identification?

Closed loop identification

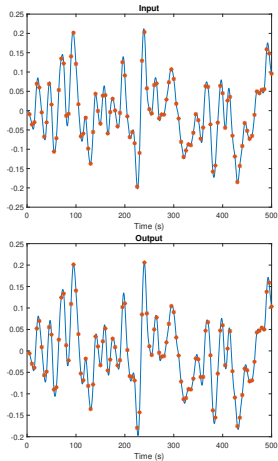
Close up



Sampling

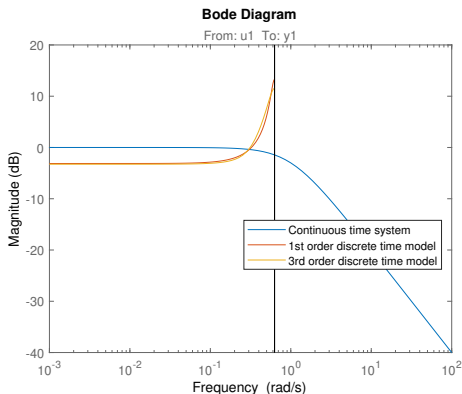
$$G(s) = \frac{1}{s + 1}$$

Data:



Sampling

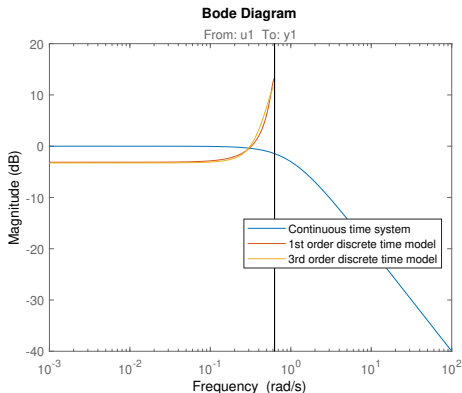
$$y(nT) = \frac{\sum_{k=1}^n b_k q^{-k}}{1 + \sum_{k=1}^n f_k q^{-k}} u(nT)$$



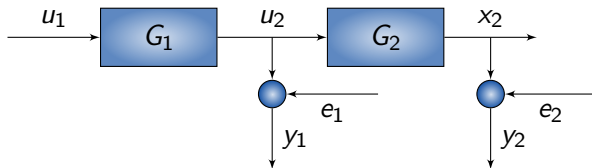
Noise free data, fast sampling. Yet problem???

Sampling

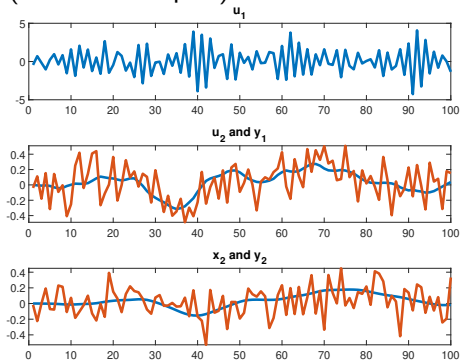
$$y(nT) = \frac{\sum_{k=1}^n b_k q^{-k}}{1 + \sum_{k=1}^n f_k q^{-k}} u(nT)$$



Measurement errors

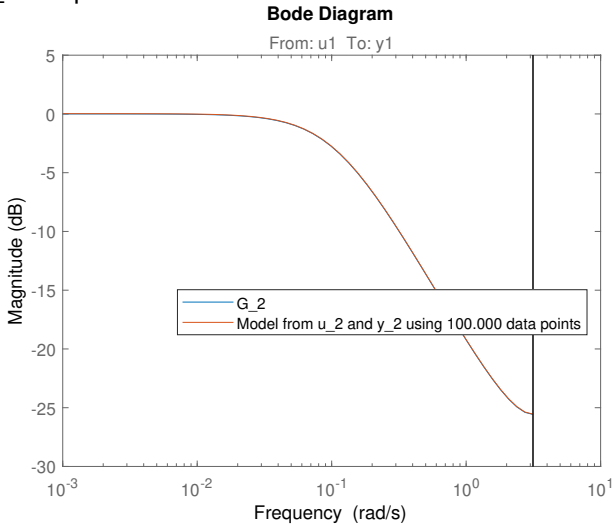


Interested in G_2 but also G_1 (high order) unknown
Large data set (100.000 samples). First 1000 shown



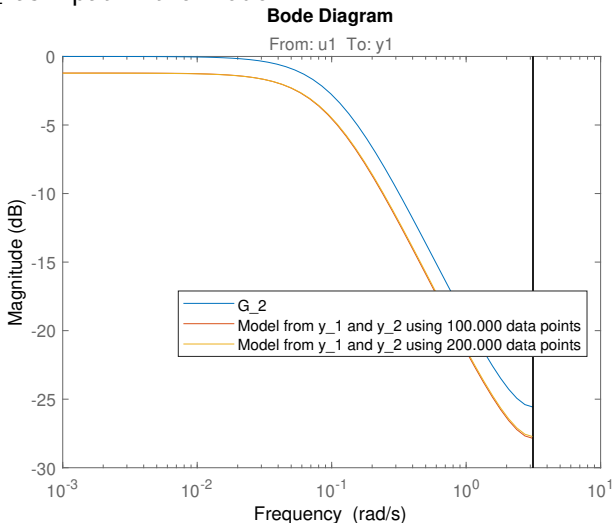
Measurement errors

Using u_2 as input



Measurement errors

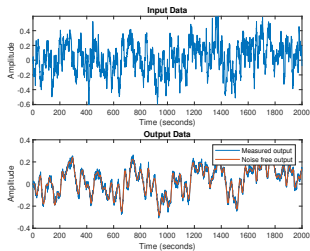
Using y_1 as input in the model



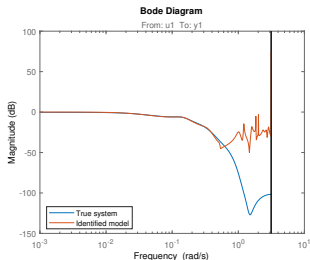
How handle measurement errors on inputs?

Complex models

System of known order 25

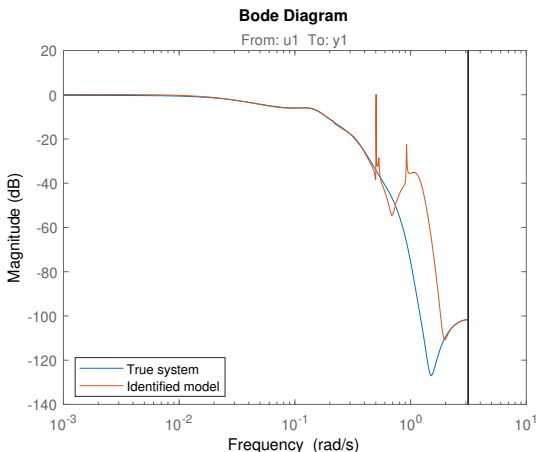


State-of-the art:



Complex models

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



Still problems. How to deal with complex systems?

Summary

Introduction

- Practicalities

- Outline

Signals

- Continuous time signals

- Discrete time signals

Dynamic systems

Introduction to parameter estimation

- Some examples

- Key problem

- Choosing the ranking function

- Summary

Inspiring pitfalls