PERSPECTIVES
on the
PROCESS
of
IDENTIFICATION

Lennart Ljung
Linköping
Questions

- How do the rudder angles affect the pitch rate?

- Aerodynamical derivatives?

- How to use the information in flight data?
Questions

- Time mark the pulp as it passes through the different vessels!
- What about the residence time in the vessels?
- How to use the information in the observed data?
The Engineer's Perspective

- How to use the information in the observed data to build a model?

- How to know if the model is any good?

- What kind of software is available for the tasks?
The Essence of the Problem

- See $z(t) = [y(t), \varphi(t)]$ for $t = 1, 2, \ldots, N$

- $\varphi(t)$: "Available Information, Past Data"

- Now see $\varphi(N + 1)$!

- Say something about $y(N + 1)$!

- $y(t)$ and $\varphi(t)$ could take values in any kind of sets.
Patterns

What we have really is a number of points in $\mathbb{R}^d, d = \dim y + \dim \varphi$.

See the pattern!
Two Basic Problems:

- Cannot have all possible $\varphi(t)$ in the observed data set.
  - Interpolation, extrapolation

- No Exact Reproducibility
  - "noise", disturbance assumptions
Perspectives

• Statistical
  – model for non-reproducibility

• Pattern Recognition
  – $y$ discrete-valued

• Projection methods (in statistics)
  – subspaces
    – linear/nonlinear regressions
• Learning theory
  — how many data points are required to distinguish patterns?

• Machine learning, Knowledge Acquisition
  — build up rules from examples
  — trees
Bottom line:

- Parameterize the "data cluster areas"!
  
  - $h(y(t), \varphi(t), \theta) \approx 0$
  
  - The function $h$ provides for the extra-and interpolations
  
  - Adjust $\theta$ using the "examples" of $\{y(t), \varphi(t)\}$

- Non-reproducibility $\iff \approx$
The Control Scientist's Perspective: System Identification

The two basic problems:

- Interpolations and extrapolations over the data-space is the task of the Model Structure

- Non-reproducibility is blamed on the Unmeasured Input $v(t)$ $\Rightarrow$ Average out by redundancy in a selection criterion.
How to cope with the unmeasured input ("disturbances, noise")?

How to pick a "selection rule"?

- Constrain the set of possible $v$:s

$$|v(t)| \leq C \quad \forall t$$

- Assign probabilities to the different possible $v$:s:

$v$ has pdf $p_v(\cdot, \theta)$
Approaches

- Non-probabilistic $v(t) \in V$
  - Unknown-but-bounded
  - Set membership

- Probabilistic
  - The pdf for $v$ gives a pdf for $z$
  - Maximum likelihood
• Pragmatic
  
  $\hat{y}(t|\theta) = g_t(\theta, \varphi(t))$  \textit{The Model Structure}
  
  $y(t) = \hat{y}(t|\theta) + e(t)$
  
  $\min V(\theta) = \sum ||y(t) - \hat{y}(t|\theta)||$

  * Contains ML and set membership
The Crux:
The Model Structure
How to extra-/interpolate over the data-space

\[ \tilde{y}(t|\theta) = g_t(\theta, \varphi(t)) \]

- Black-Box
- Physical Modeling
- Semi-Physical Modeling
The Crux: 
The Model Structure
How to extra-/interpolate over the data-space

\[ \hat{y}(t|\theta) = g_t(\theta, \varphi(t)) \]

- Black-Box
- Physical Modeling
- Semi-Physical Modeling
Black Boxes

Idea: Interpolate between the $\varphi$:s by smooth standard functions

$$\hat{y}(t|\theta) = \sum_{k=1}^{d} \theta_k h_k(\varphi(t))$$

$\varphi(t) = [y(t-1), \ldots, y(t-n), u(t-1), \ldots, u(t-m)]$

$h_k(\varphi)$ are basis functions that are mappings from the $\varphi$-space to the $y$-space. They may depend on $\theta$:

$h_k(\varphi, \theta)$
Black-Box Basis Functions

Basic property:

- \( h_k(\varphi), k = 1, \ldots \) form a basis for all (reasonable) functions from the \( \varphi \)-space to the \( y \)-space.

- \( d = d(N) \rightarrow \infty \) as \( N \rightarrow \infty \): Non-parametric (regression) methods.

- Hope to “do well” with just a few of them
Character of the basis functions:

- **Local**

- **Global**
Common Choices of Basis Function

- "Classic System Identification"
  - Linear $\varphi$-spaces: $h_k(\varphi(t)) = u(t - k)$ or $y(t - k)$ (or $\hat{y}(t - k|\theta)$): The black-box difference equation family. (ARX, ARMAX, etc)

  Can also be viewed as bases in the space of frequency functions:

  ![Graph](image)

  - Volterra and other non-linear counterparts
• "Classic non-parametric regression"
  
  - Nearest Neighbor: $h_k(\varphi)$ indicator function for smallest possible data box

  - Average boxes (Radial basis Neural Networks): (Smooth) indicator function for somewhat bigger boxes.

  - Trees
• Neural Networks

  — Explicit equations for $h_k$ complicated, but easy recursions

• Fuzzy Models

  — Membership functions — interpolation
  functions — $h_k$
Physical Model Structures

Basic Guideline: Don't Estimate What You Already Know!

The Physics is used to interpolate and EXTRAPOLATE in the ϕ-space
Semi-Physical Model Structures

Introduce essential non-linearities "by hand"

Again: Don't estimate what you already know

\[ r_1(t) \rightarrow u_1(t) = r_1^2(t) - m_2(t) \]
\[ m_2(t) \rightarrow \text{THINK!} \rightarrow \hat{y}(t|\theta) \]
\[ w(t) \rightarrow y(t) = w(t) \cdot r_1(t) \]
The Heart of the Matter: Model Validation

The basic process of identification can be seen as a way to provide candidate models to be subjected to validation:

- How far away might it be from a correct description?
  - Next page!

- Are my model structure assumptions consistent with the observed data?
  - (Classical) residual analysis

- Is it good enough?
  - Subjective!
“Model Error Modeling”

- Again the two basic problems:
  - Not the right interpolation rules; Bias Error
  - Getting fooled by the “noise”; Random Error
• Basic Advice:
  – Determine a model that passes the validation tests.
  – $\Rightarrow$ Bias error $\leq$ random error
  – Reduce model if necessary – with respect to its purpose
The Engineer’s Perspective II
Solving the Problem

A recipe for dynamical systems:

1. compare(z,arx(z(1:200,:),[4 4 1]))

2. Does it look good?
   - Yes: Congratulations!
   - NO:
     - Higher order
     - More inputs
     - Apply semi-physical modelling
     - Give up!
Aircraft Dynamics

Dashed line: Actual Pitch rate. Solid line: 10 step ahead predicted pitch rate, based on the fourth order model from canard angle only.
As above but using all three inputs.
Buffer Vessel Dynamics

Dashed line: $\kappa$-number after the vessel, actual measurements. Solid line: Simulated $\kappa$-number using the input only and a fourth order linear model with delay 12, estimated using the first 200 data points.

Think: ....
z = [y,u]; pf = flow./level;
t = 1:length(z)
newt = table1([cumsum(pf),t],[pf(1):sum(pf)]');
newz = table1([t,z],newt);

Same as previous figure but applied to resampled data
What's the impulse response of our model?

\[ m = \text{arx}(\text{ze}, \text{nn}); \]
\[ \text{impres} = \text{idsim}([1; \text{zeros}(49, 1)], m); \]
\[ \text{plot(impres)} \]
What's the uncertainty?

\texttt{idssimsd([1;zeros(49,1)],m)}
Conclusions

- Process identification is meeting place for practical problems and fairly advanced theory
- The pragmatic approach ("Curve fitting") has many theoretical interpretations
- Important to see the links between "hot" new approaches and classic theory
- Good software support
- The area starts and ends with real data
Bottom line:
See the pattern in observed data!