Challenges of Non-linear System Identification

Lennart Ljung
Linköping University

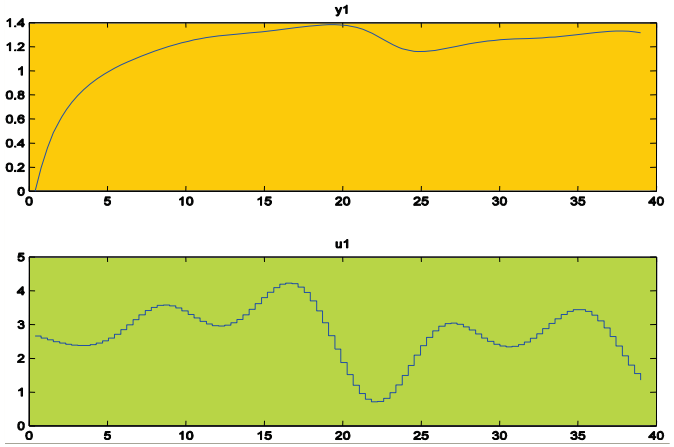
Prologue

Prologue

The PI, the Customer and the Data Set

- C: I have this data set. I have collected it from a cell metabolism experiment. The input is Glucose concentration and the output is the concentration of G6P. Can you help me building a model of this system?

The Data Set



Output

Input

A Simple Linear Model

Try the simplest model

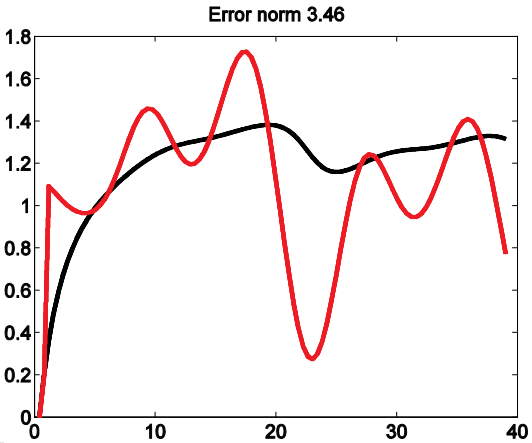
y(t) = a u(t-1) + b u(t-2)

Fit by Least Squares:

m1=arx(z,[0 2 1])

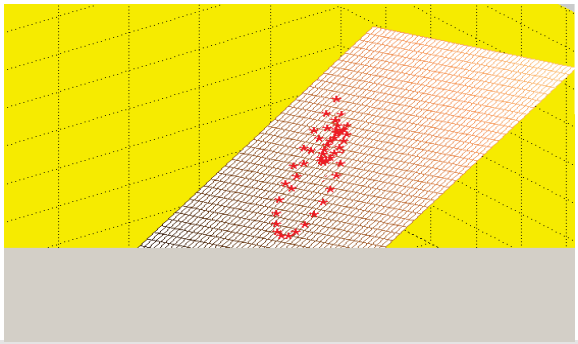
compare(z,m1)

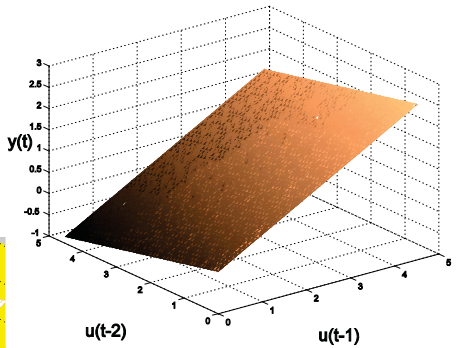
Red: Model
Black: Measured



A Picture of the Model

Depict the model as y(t) as a function of u(t-1) and u(t-2)



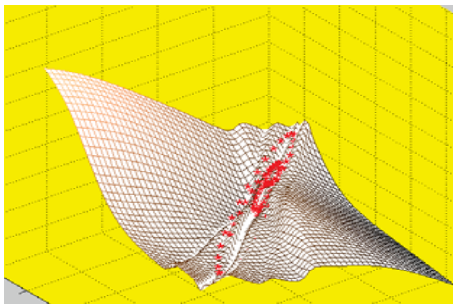


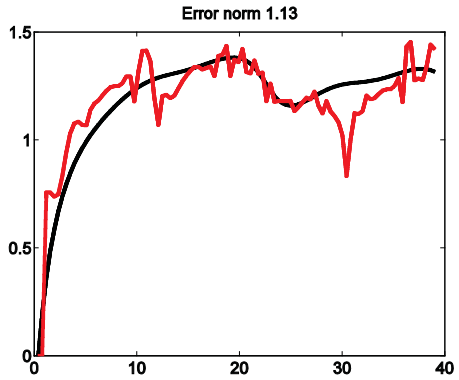
A Nonlinear Model

Try a nonlinear model

y(t) = f(u(t-1),u(t-2))

```
m2 = arxnl(z,[0 2 1],'sigm')
compare(z,m2)
```

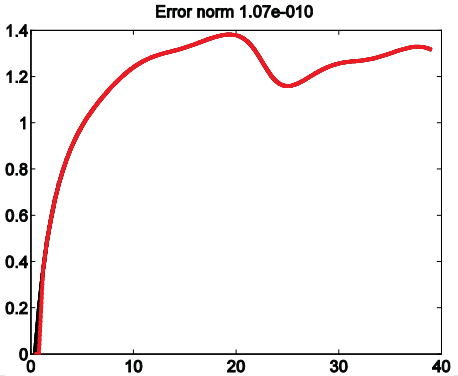


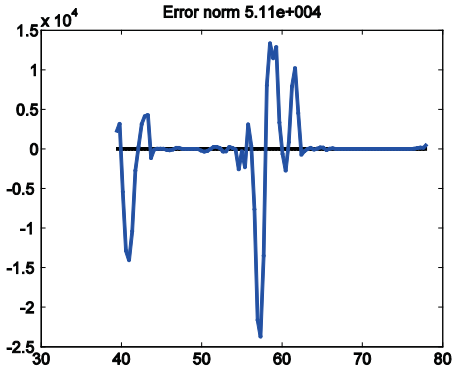


More Flexibility

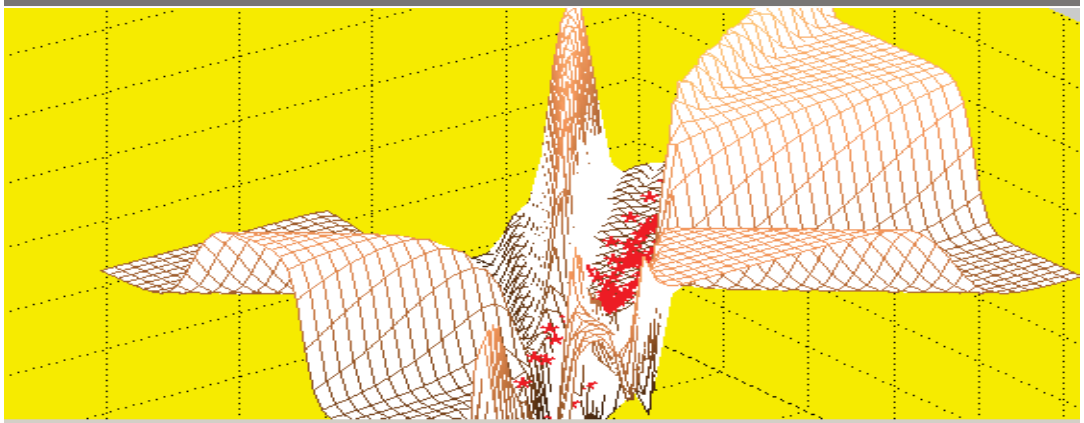
A more flexible, nonlinear model

```
y(t) = f(u(t-1),u(t-2))
m3 = arxnl(z,[0 2 1],'sigm','numb',100)
compare(z,m3)
compare(zv,m3)
```



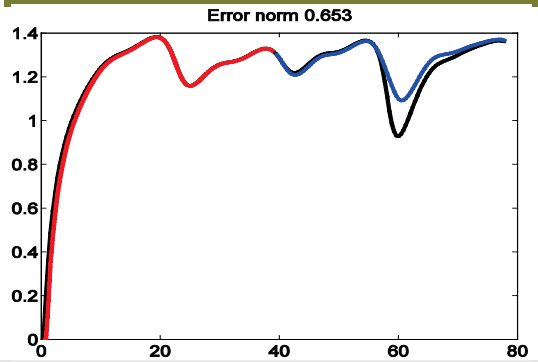


The Fit Between Model and Data



More Regressors

Try other arguments:
y(t) = f(y(t-1),y(t-2),u(t-1),u(t-2))
m4 = arxnl(z,[2 2 1],'sigm')
compare([z;zv],m4)



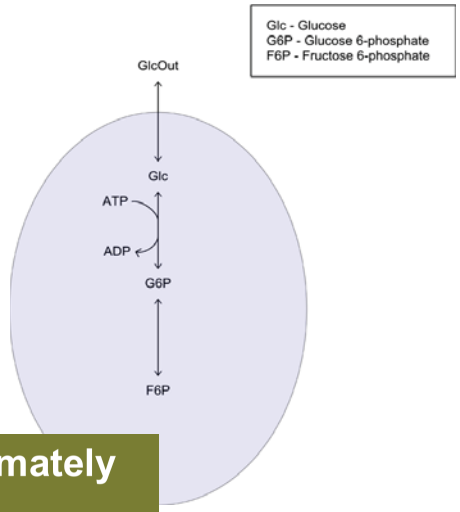
Biological Insight

Pathway diagram

$$\dot{x}_1 = -\theta_1 \frac{x_1/\theta_2 - x_2/\theta_3}{1 + x_1/\theta_2 + x_2/\theta_3} + \theta_4 \frac{u - x_1}{1 + u/\theta_5 + x_1/\theta_5 + x_1 u/\theta_5^2}$$

$$\dot{x}_2 = \theta_1 \frac{x_1/\theta_2 - x_2/\theta_3}{1 + x_1/\theta_2 + x_2/\theta_3} - \theta_6 \frac{x_2/\theta_7 - \theta_8}{1 + x_2/\theta_7 + \theta_8}$$

$$y = x_2$$



For sampled data, approximately

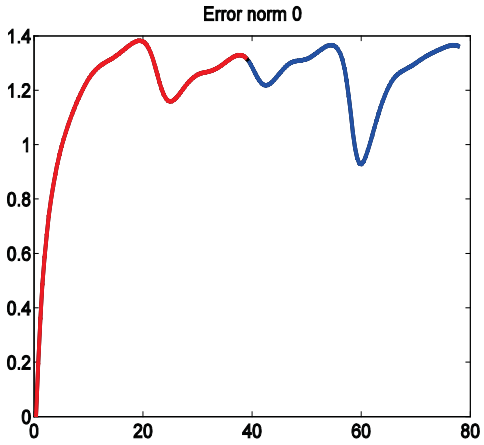
y(t) = f(y(t-1),y(t-2),u(t-1),u(t-2),θ)

Tailor-made Model Structure

cell = nlgrey(eqns,nom_pars)

m5 = pem(z,cell);

compare([z;zv],m5)



End of Prologue

Outline

- Problem formulation
- How to parameterize black box predictors
- What are the ”optimal choices”?
- Sparsity
- Using physical insight
- Initialization of parameter search
- How to live with LTI approximations
- www.control.isy.liu.se/~ljung/bode

The Basic Picture

Input u, Output y, $\quad Z^t = \{u(1), y(1), \ldots u(t), y(t)\}$

- State-Space

$$\dot{x} = g(x, u, w)$$
$$y = h(x, u, e)$$
w and e noises

$$\hat{y}(t|t-1) = E(y(t)|Z^{t-1})$$

- Output predictor

$$\hat{y}(t|t-1) = f_0(Z^{t-1})$$

The Predictor Function

General structure $\quad \hat{y}(t|t-1) = f_0(Z^{t-1})$

Common/useful special case:

$$\hat{y}(t|t-1) = f_0(Z^{t-1}) = f_0(\phi(t))$$
$\phi(t) = \phi(Z^{t-1})$ of fixed dimension m (”state”, ”regressors”)

Think of the simple case

$$\phi(t) = \begin{bmatrix} y(t-1) & \ldots & y(t-n_a) & u(t-1) & \ldots & u(t-n_b) \end{bmatrix}$$

The Predictor Function

General structure $\quad \hat{y}(t|t-1) = f_0(Z^{t-1})$

Common/useful special case:

$$\hat{y}(t|t-1) = f_0(Z^{t-1}) = f_0(\phi(t))$$
$\phi(t) = \phi(Z^{t-1})$ of fixed dimension m (”state”, ”regressors”)

Think of the simple case

$$\phi(t) = \begin{bmatrix} y(t-1) & \ldots & y(t-n_a) & u(t-1) & \ldots & u(t-n_b) \end{bmatrix}$$

- IMPORTANT PROBLEM
- Get inspiration from state space model to select ϕ (nonlinear observer)
-

Another Time!

The Data and the Identification Process

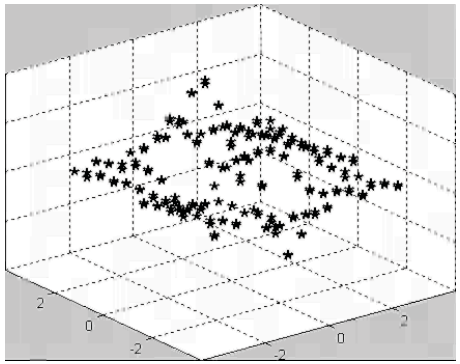
The observed data
$Z^N=[y(1),\phi(1),\ldots y(N),\phi(N)]$

are N points in $\mathbb{R}^{m+1}$

The predictor model

$$\hat{y} = f_0(\phi)$$

is a surface in this space



Identification is to find the predictor surface from the data:

Mathematical Formulation

- Parameterize the predictor function: f(θ,ϕ), f∈$\mathcal{F}$ when θ ∈ D
- Collect observations: Z^N
- Fit the parameters to the data

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

$$V_N(\theta, Z^N) = \sum_{t=1}^{N} \ell(y(t) - f(\theta, \phi(t)))$$

$$= \sum_{t=1}^{N} \|y(t) - f(\theta, \phi(t))\|^2$$

- Use model $\hat{f}_N(\phi) = f(\hat{\theta}_N, \phi)$

Mathematical Formulation

- Parameterize the predictor function: f(θ,ϕ), f∈$\mathcal{F}$ when θ ∈ D
- Collect observations: Z^N
- Fit the parameters to the data

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

$$V_N(\theta, Z^N) = \sum_{t=1}^{N} \ell(y(t) - f(\theta, \phi(t)))$$

- IMPORTANT PROP
- The fit for esti data $V_N(\hat{\theta}_N, Z^N)$ is known
- How to the fit for another (validation) data set?

Another Time!

- Use model $\hat{f}_N(\phi) = f(\hat{\theta}_N, \phi)$

Outline

- Problem formulation
- How to parameterize black box predictors
- What are the ”optimal choices”?
- Sparsity
- Using physical insight
- Initialization of parameter search
- How to live with LTI approximations
- www.control.isy.liu.se/~ljung/bode

The First Challenge

❖How to parameterize the predictor function f(θ,ϕ)?

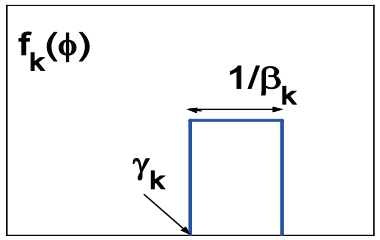
- ➢ Grey-box (Physical insight of some sort)
- ➢ Black-box (Flexible function expansions)

$$f(\theta,\phi)=\sum_{k=1}^{n}\theta_k f_k(\phi)$$

General case: $f_k(\phi)=f_k(\phi(\theta),\theta)$

Choice of Functions: Methods

- Neural Networks
- Radial Basis Neural Networks
- Wavelet-networks
- Neuro-Fuzzy models
- Spline networks
- Support Vector Machines
- Gaussian Processes
- Kriging



$\gamma_k=\phi(k)$

$\kappa=\text{GaussianBell}$

ALL THESE USE $f(\theta,\phi)=\sum_{k=1}^{n}\alpha_k\kappa(\beta_k(\phi-\gamma_k))$

Several layers…. $\theta=\{\alpha_k,\beta_k,\gamma_k\}$

The Second Challenge

- What is the best choice of regressor basis functions

$$f_k \text{ in } \textstyle\sum_{k=1}^{d}\theta_k f_k(\phi)?$$

- What choice of f_k minimizes the mean square error between the true $f_0(\phi^*)$ and the estimate $\hat{f}_N(\phi^*)$ at a particular value ϕ^* ?

- Ask a different question: Not so special ! $f_k \leftrightarrow c_t$
- Given ϕ^* and past observations {y(t),ϕ(t),t=1,…N}, what would be the best choice of weights c_t in an estimate

$$\hat{f}_N(\phi^*)=\sum_{t=1}^{N}c_t y(t)$$

DWO: More detailed calculations

- Assume that $y(t)=f_0(\phi(t))+e(t);\quad Ee^2(t)=\sigma^2$
- Let the mean square error at the point ϕ^* be

$$H(f_0,c)=E(f_0(\phi^*)-\hat{f}_N(\phi^*))^2=$$
$$\textstyle\sum_{t=1}^{N}c_t^2\sigma^2+\left(\sum_{t=1}^{N}c_t(f_0(\phi(t))-f_0(\phi^*))+(1-\sum_{t=1}^{N}c_t)f_0(\phi^*)\right)^2$$

Assume that we know that $f_0\in\mathcal{F}$, some class of functions.

Then a natural choice of c is $\min_{c_t}\max_{f_0\in\mathcal{F}}H(f_0,c)$

DWO: Direct Weight Optimization

If $\mathcal{F}$ contains unbounded functions, we must require $\sum c_t = 1$

The Optimization

$$\min_{c_t} \max_{f_0 \in \mathcal{F}} \sum_{t=1}^{N} c_t^2 \sigma^2 + \left(\sum_{t=1}^{N} c_t (f_0(\phi(t)) - f_0(\phi^*)) \right)^2$$

This is a convex minimization problem!

What about $\mathcal{F}$?

Theorem: $\mathcal{F} = \{f | f(\theta, \phi) = \sum_{k=1}^{d} \theta_k f_k(\phi) \text{ for some } \theta_k\}$.

$$c_t^{opt} = F^T(\phi(k)) \left[\sum_{k=1}^{N} F(\phi(k)) F^T(\phi(k)) \right]^{-1} F(\phi^*)$$

$$F^T(\phi(k)) = \left[f_1(\phi(k)), \ldots, f_d(\phi(k)) \right]$$

The LS solution!

Another Class of Functions

$$\min_{c_t} \max_{f_0 \in \mathcal{F}} \sum_{t=1}^{N} c_t^2 \sigma^2 + \left(\sum_{t=1}^{N} c_t (f_0(\phi(t)) - f_0(\phi^*)) \right)^2$$

Suppose $\mathcal{F} = \{f | f \text{ differentiable and } |f'(x_1) - f'(x_2)| \leq L|x_1 - x_2|\}$

Then

$$H(f_0, c) \leq \sum_{t=1}^{N} c_t^2 \sigma^2 + \frac{L^2}{4} \left(\sum_{t=1}^{N} \tilde{\phi}^2(t) |c_t| \right)^2$$

$$\tilde{\phi} = \phi - \phi^*$$

Minimizing the RHS is a QP

The structure means that "many" c_t are zero

$\mathcal{F}$ is a Set of Differentiable Functions

- The DWO solution is (Roll, Nazin, Ljung, 2002)

$$\hat{f}_N(\phi^*) = \sum_{t=1}^{N} c_t y(t)$$

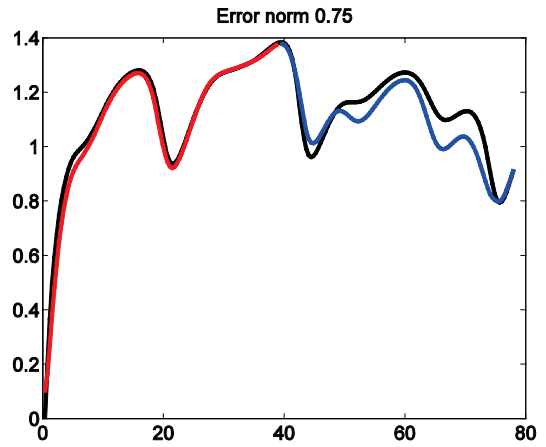
$$\begin{cases} c_t = 0 \text{ if } |\phi(t) - \phi^*| > C(L/\sigma) \\ c_t \text{ on one or two parabola pieces otherwise} \end{cases}$$

- Note: Finite "Bandwidth": The estimate depends only on observations in the vicinity of the target point. The choice of bandwidth is automatic, given L/σ

Test this Approach on the Cell Data

Test it on the cell data: Compare([z;zv],mdwo)

L and σ have been estimated from data

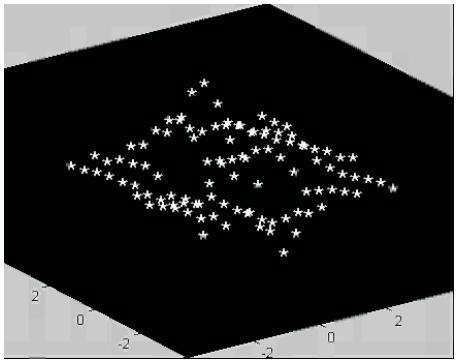


Challenges Dealt with so Far

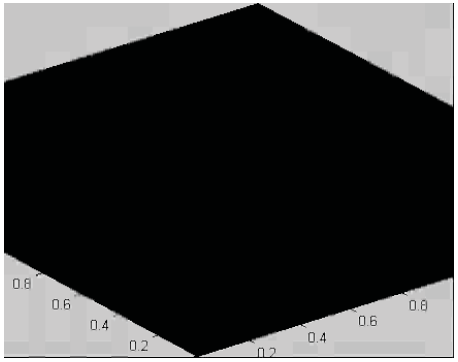
- How can the black-box predictor model be parameterized
- What is the ”best” black box approach?
- Now comes the real challenge …

The Real Challenge: Sparsity

2 regressors 100 points



10 regressors 10^7 points (unit cube, 2D proj from random points, visibility 0.1)



Outline

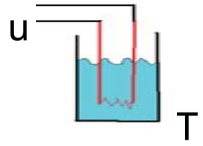
- Problem formulation
- How to parameterize black box predictors
- What are the ”optimal choices”?
- Sparsity
- Using physical insight
- Initialization of parameter search
- How to live with LTI approximations
- www.control.isy.liu.se/~ljung/bode

How to Deal with Sparsity

- Need ways to interpolate and extrapolate in the data space
- Leap of Faith: Search for global patterns in observed data to allow for data-driven interpolation
- Use Physical Insight: Allow for few parameters to parameterize the predictor surface, despite the high dimension.

Using Physical Insight I

Semiphysical Modeling



Input: heater voltage u

Output: Fluid temperature T

Square the voltage: $u \rightarrow u^2$

Using Physical Insight I

Semiphysical Modeling

Hammerstein-Wiener

Using Physical Insight I

Semiphysical Modeling

Hammerstein-Wiener

Local Linear Models (also LPV)

$\phi = [\rho, \psi]$; $f(\theta,\phi)=f(\theta,\rho,\psi)$

Linear in ψ for fixed ρ (”regime variable”)

Using Physical Insight II

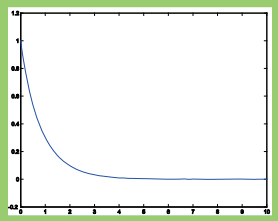
- Careful modeling leading to systems of Differential Algebraic Equations (DAE) parameterized by physical parameters.
- Support by modern modeling tools.
- The ”statistically correct” approach is to estimate the parameters by the Maximum Likelihood method.

Maximum Likelihood: The Solution?

- Example: A Michaelis-Menten equation:

$$\dot{x} = \theta_1 \frac{x}{\theta_2 + x} - x + u$$
$$y = x + e$$
$$u = \text{impulse}$$

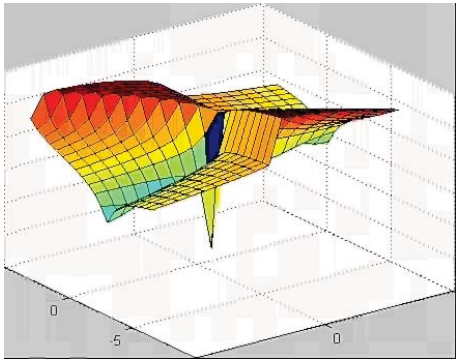
- The output:



The ML Criterion (Gaussian Noise)

V(θ) as a function of θ

$$V(\theta) = \sum_{k=1}^{100} \|y(t_k) - x(t_k, \theta)\|^2$$
$$\dot{x}(t,\theta) = \theta_1 \frac{x(t,\theta)}{\theta_2 + x(t,\theta)} - x(t,\theta) + u(t)$$



Outline

- Problem formulation
- How to parameterize black box predictors
- What are the ”optimal choices”?
- Sparsity
- Using physical insight
- Initialization of parameter search
- How to live with LTI approximations
- www.control.isy.liu.se/~ljung/bode

Identifiability and Linear Regression

Crucial Challenge for physically parameterized models: Find a good initial estimate

- Result of conceptual interest: (Ljung, Glad, 1994)

A parameterized set of DAEs is globally identifiable

if and only if

the set can be rearranged as a linear regression

Ritt’s algorithm for differential algebra provides a finite procedure for constructing the linear regression

A More Formal Description

- The set of DAE's: $g_i(y,u,x,\theta,d/dt)=0$; i=1,…,r
- Any signals that satisfy all the equations above will also satisfy any equation obtained by differentiating, adding, and multiplying equations in this set.
- This gives an infinite amount of equations G.
- Find a subset (characteristic set) of G that has the same solution set and is as "simple as possible"
- Simple: No x and no high exponents of θ
- Ritts algorithm makes the construction of such a set
- The set of DEAs is globally identifiable if and only if the set contains an equation $\psi(y,u,d/dt) = \theta\ \rho(y,u,d/dt)$

Example of Ritt's Algorithm

Original equations

$$\dot{x}_1 = \theta x_2^2$$
$$\dot{x}_2 = u$$
$$y = x_1$$

Differentiate y twice

$$\dot{y} = \dot{x}_1 = \theta x_2^2$$
$$\ddot{y} = 2\theta x_2 \dot{x}_2 = 2\theta x_2 u$$

Square the last expression

$$\ddot{y}^2 = 4\theta\theta x_2^2 u^2 = 4\theta \dot{y} u$$

which is a linear regression

The Michaelis-Menten Equation

- In our case (noisefree)

$$\dot{y} = \frac{\theta_1 y}{y+\theta_2} - y + u$$

$$\dot{y}y + \theta_2\dot{y} = \theta_1 y - y^2 - \theta_2 y + uy + \theta_2 u$$

$$\dot{y}y + y^2 - uy = \begin{bmatrix}\theta_1 & \theta_2\end{bmatrix}\begin{bmatrix} y \\ u - \dot{y} - y\end{bmatrix}$$

With noisy observations $y=x_1+e$, the noise structure in the linear regression may be violated, giving biased estimates. In this case, sufficiently good initial estimates are obtained.

Challenge for Grey Box Models

- Only small examples treated so far. Make the initialization work in bigger problems.
- Potential for important contributions:
  - Handle the complexity by modularization
  - Handle the noise corruption so that good quality initial estimates are secured
- Room for innovative ideas using algebra and semidefinite programming!

A Final Challenge

- Despite all the work and results on non-linear models, the most common situation will still be

How to live with an estimated LTI model approximation of a Non-linear system.

Non-linear System Approximation

- Given an LTI Output-error model structure y=G(q,θ)u+e, what will the resulting model be for a non-linear system?
- Assume that the inputs and outputs u and y are such that the spectra Φ_u and Φ_{yu} are well defined.
- Then the LTI second order equivalent is

$$G_0 = \frac{1}{\lambda L(z)}\left[\frac{\Phi_{yu}(z)}{L(z^{-1})}\right]_{\text{causal}} \qquad \Phi_u(z) = \lambda L(z) L(z^{-1})$$

Note: G_0 depends on u

- The limit model will be

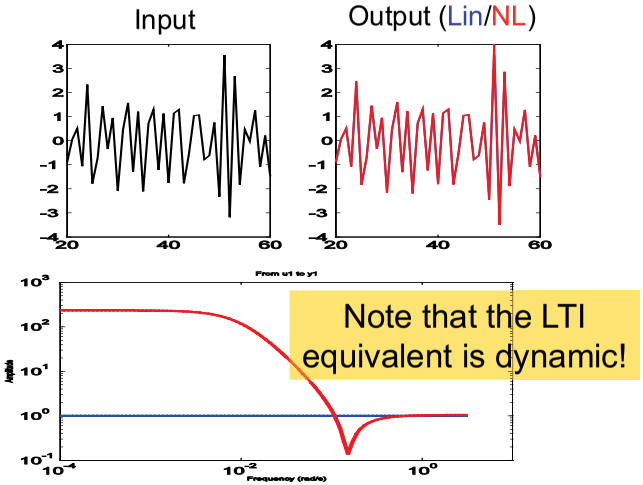
$$\min_\theta \int |G(z,\theta) - G_0(z)|^2 \Phi_u(z) dz$$

An Example

- Two data sets
- Input u and output y
- y = u
- $y = u + 0.01u^3$

(Enqvist, 2003)

The corresponding LTI equivalents (amplitude Bode plot)



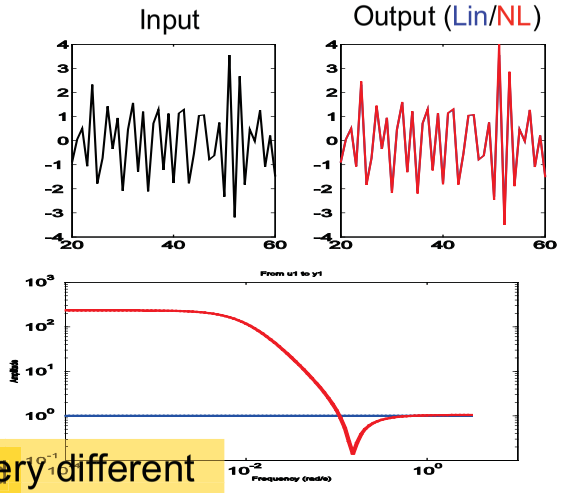
Note that the LTI equivalent is dynamic!

An Example

- Two data sets
- Input u and output y
- y = u
- $y = u + 0.01u^3$

(Enqvist, 2003)

The corresponding LTI equivalents (amplitude Bode plot)



So $oe(z, [2\ 2\ 1])$ give very different results for the two data sets!

Epilogue

- Four Challenges for the Control Community:

1) A working theory for stability of black-box models.
   - ✓ Prediction/Simulation
2) Fully integrated software for modeling and identification
   - ✓ Object oriented modeling
   - ✓ Differential algebraic equations
   - ✓ Full support of disturbance models
3) Robust parameter initialization techniques
   - ✓ Algebraic/Numeric
4) Dealing with LTI-equivalents for good control design

Thanks to …

Coauthors (non-linear identification):

Alberto Bemporad * Albert Benveniste * Martin Braun * Torbjörn Crona * Bernard Delyon * Martin Enqvist * P-Y Glorennec * Markus Gerdin * Torkel Glad * Fredrik Gustafsson * Håkan Hjalmarsson * Anatoli Juditsky * Ingela Lind * David Lindgren * Peter Lindskog * Mille Millnert * Alexander Nazin * Alexander Poznyak * Pablo Parrilo * Dan Rivera * Jacob Roll * Jonas Sjöberg * Anders Skeppstedt * Anders Stenman * Jan-Erik Strömberg * Vincent Verdult * Michel Verhaegen * Qinghua Zhang *

Help with presentation:

Jan Willems, Mats Jirstrand, Johan Gunnarsson, Jacob Roll, Martin Enquist, Rik Pintelon, Johan Schoukens, Michel Gevers, Bart deMoor, …

www.control.isy.liu.se/~ljung/bode

Lennart Ljung
Challenges of Non-linear Identification
Bode Lecture
CDC 2003
AUTOMATIC CONTROL
COMMUNICATION SYSTEMS
LINKÖPINGS UNIVERSITET