System Identification: From Data to Model

Lennart Ljung

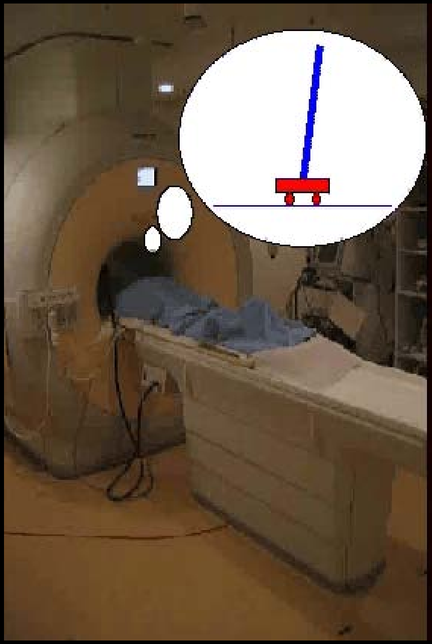
Linköping University, Sweden

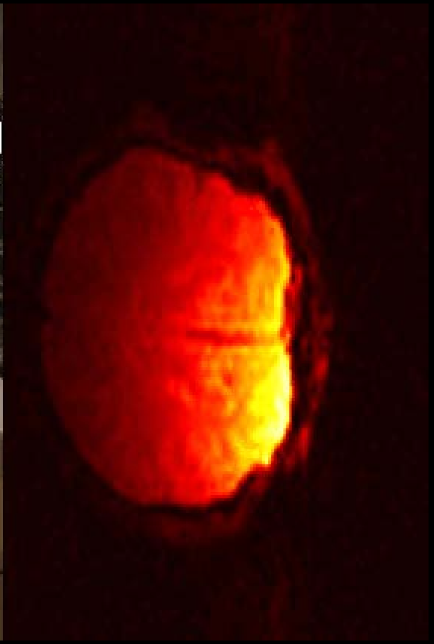
- Peter Sagirow Seminar, Stuttgart, Nov 7, 2011

The Problem

Flight tests with
Gripen at high alpha

Person in Magnet camera,
stabilizing a pendulum by
thinking ”right”-”left”





fMRI picture of brain

The Confusion

Support Vector Machines * Manifold learning *prediction error method * Partial Least Squares * Regularization * Local Linear Models * Neural Networks * Bayes method * Maximum Likelihood * Akaike's Criterion * The Frisch Scheme * MDL * Errors In Variables * MOESP * Realization Theory *Closed Loop Identification * Cram\'er - Rao * Identification for Control * N4SID* Experiment Design * Fisher Information * Local Linear Models * Kullback-Liebler Distance * MaximumEntropy * Subspace Methods * Kriging * Gaussian Processes * Ho-Kalman * Self Organizing maps * Quinlan's algorithm * Local Polynomial Models * Direct WeightOptimization * PCA * Canonical Correlations * RKHS * Cross Validation *co-integration * GARCH * Box-Jenkins * Output Error * Total Least Squares * ARMAX * Time Series * ARX * Nearest neighbors * Vector Quantization *VC-dimension * Rademacher averages * Manifold Learning * Local Linear Embedding* Linear Parameter Varying Models * Kernel smoothing * Mercer's Conditions *The Kernel trick * ETFE * Blackman--Tukey * GMDH * Wavelet Transform * Regression Trees * Yule-Walker equations * Inductive Logic Programming *Machine Learning * Perceptron * Backpropagation * Threshold Logic *LS-SVM * Generaliztion * CCA * M-estimator * Boosting * Additive Trees * MART * MARS * EM algorithm * MCMC * Particle Filters *PRIM * BIC * Innovations form * AdaBoost * ICA * LDA * Bootstrap * Separating Hyperplanes * Shrinkage * Factor Analysis * ANOVA * Multivariate Analysis * Missing Data * Density Estimation * PEM *

This Talk

Two objectives:

- Place System Identification on the global map. Who are our neighbours in this part of the universe?
- Discuss some open areas in System Identification.

The Communities

- Constructing (mathe-matical) models from data is a prime problem in many scientific fields and many application areas.
- Many communities and cultures around the area have grown, with their own nomenclatures and their own ``social lives".
- This has created a very rich, and somewhat confusing, plethora of methods and approaches for the problem.

.

A picture: There is a core of central material, encircled by the different communities

The Core

Model $\mathfrak{m}$ – Model Set $\mathcal{M}$ – Complexity (Flexibility) $\mathcal{C}$

Information $\mathcal{I}$ – Data Z

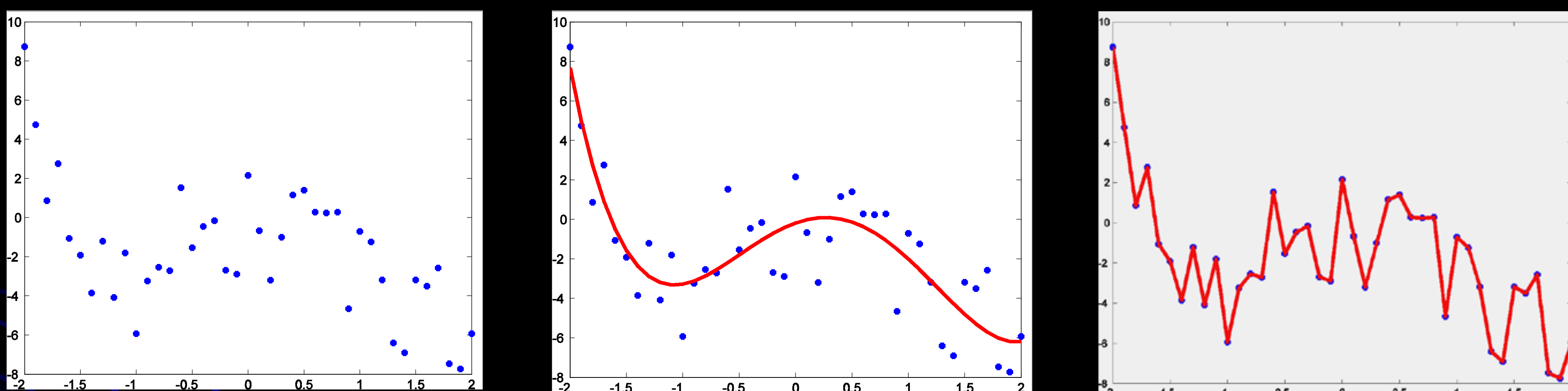
Estimation – Validation (Learning – Generalization)

Model fit $\mathcal{F}(\mathfrak{m}, Z)$

Estimation

Squeeze out the relevant information in data

But NOT MORE !

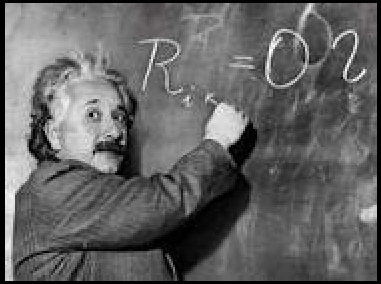


All data contain information and misinformation ("Signal and noise")

So need to meet the data with a prejudice!

Estimation Prejudices

- Nature is Simple!
  - Occam's razor



  - God is subtle, but He is not malicious (Einstein)

- So, conceptually:

$$\hat{\mathfrak{m}} = \arg\min_{\mathfrak{m}\in\mathcal{M}} (\text{Fit} + \text{Complexity Penalty})$$

Estimation and Validation

Fit to estimation data Z_e^N (N: Number of data points)

$$F(\hat{\mathfrak{m}}, Z_e^N) \qquad \text{("The empirical risk")}$$

Now try your model on a fresh data set (Validation data Z_v):

$$E\mathcal{F}(\hat{\mathfrak{m}}, Z_v) \approx \mathcal{F}(\hat{\mathfrak{m}}, Z_e^N) + f(\mathcal{C}(\mathcal{M}), N)$$

f is a function of the complexity, so the more flexible the model set the more the expected fit to validation data is deteriorated. (Exact formulations: Akaike's FPE (AIC), Vapnik's learning/generalization result, Rademacher averages ...)

So don't be impressed by a good fit to estimation data in a flexible model set!

Bias and Variance

$\mathcal{S}$ – True system $\quad \hat{\mathfrak{m}}$ – Estimate $\quad \mathfrak{m}^* = E\hat{\mathfrak{m}}$

$\hat{\mathfrak{m}} \in \mathcal{M}$: Typically $\mathfrak{m}^*$ is the model closest to $\mathcal{S}$ in $\mathcal{M}$.

$$E\|\mathcal{S} - \hat{\mathfrak{m}}\|^2 = \|\mathcal{S} - \mathfrak{m}^*\|^2 + E\|\hat{\mathfrak{m}} - \mathfrak{m}^*\|^2$$

MSE	=	BIAS (B)	+	VARIANCE (V)
Error	=	Systematic	+	Random

As $\mathcal{C}(\mathcal{M})$ increases, B decreases &V increases

This bias/variance tradeoff is at the heart of estimation!

Note that the $\mathcal{C}$ that minimizes the MSE typically has a B$\neq 0$!

Information Contents in Data and the CR Inequality

The value of information in data depends on prior knowledge. Observe Y. Let its probability density function be $f_Y(x,\theta)$ The (Fisher) Information Matrix is

$$\mathcal{I} = E\ell'_Y(\ell'_Y)^T, \qquad \ell'_Y = \frac{\partial}{\partial\theta}\log f_Y(x,\theta)$$

The Cramér-Rao inequality tells us that

$$\mathrm{cov}\hat{\theta} \geq \mathcal{I}^{-1}$$

for any (unbiased) estimator $\hat{\theta}$ of the parameter.

$\mathcal{I}$ is thus a prime quantity for Experiment Design.

The Communities Around the Core I

- Statistics : The mother area
  - … EM algorithm for ML estimation
  - Resampling techniques (bootstrap…)
  - Regularization: LARS, Lasso …
- Statistical learning theory
  - Convex formulations, SVM (support vector machines)
  - VC-dimensions
- Machine learning
  - Grown out of artificial intelligence: Logical trees, Self-organizing maps.
  - More and more influence from statistics: Gaussian Processes, HMM (Hidden Markov Models), Baysian nets

The Communities Around the Core II

- Manifold learning
  - Observed data belongs to a high-dimensional space
  - The action takes place on a lower dimensional manifold: Find that!
- Chemometrics
  - High-dimensional data spaces
    (Many process variables)
  - Find linear low dimensional
    subspaces that capture the essential state: PCA, PLS (Partial Least Squares), ..
- Econometrics
  - Volatility Clustering
  - Common roots for variations

The Communities Around the Core III

- Data mining
  - Sort through large data bases looking for information: ANN, NN, Trees, SVD…
  - Google, Business, Finance…
- Artificial neural networks
  - Origin: Rosenblatt's perceptron
  - Flexible parametrization of hyper-surfaces
- Fitting ODE coefficients to data
  - No statistical framework: Just link ODE/DAE solvers to optimizers
- System Identification
  - Experiment design
  - Dualities between time- and frequency domains

System Identification – Past and Present

Two basic avenues, both laid out in the 1960's

- Statistical route: ML etc: Åström-Bohlin 1965
  - Prediction error framework: postulate predictor and apply curve-fitting
- Realization based techniques: Ho-Kalman 1966
  - Construct/estimate states from data and apply LS (Subspace methods).

Past and Present:

- Useful model structures
- Adapt and adopt core’s fundamentals
- Experiment Design ….
  - ...with intended model use in mind (”identification for control”)

Example: Aircraft Dynamics

Five inputs and two outputs.
Build models of the kind

$$x(t+1) = Ax(t) + Bu(t) + Ke(t)$$
$$y(t) = Cx(t) + e(t)$$

"order" $=$ dim x.

Inputs

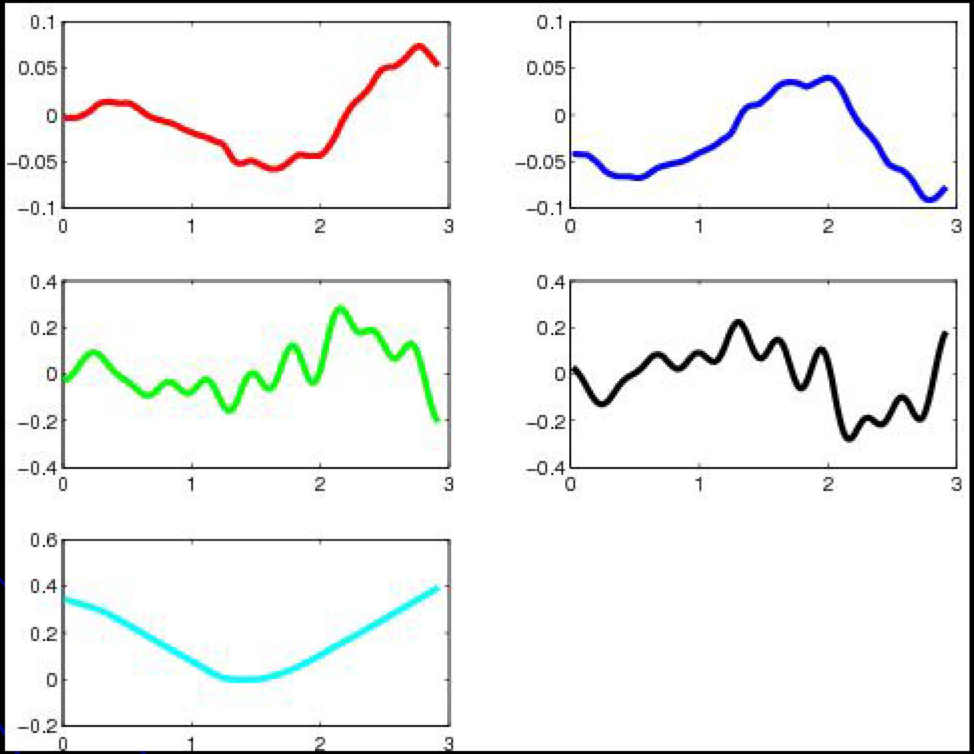
Elevator

Canard

Deriv. Elev.

Deriv. canard

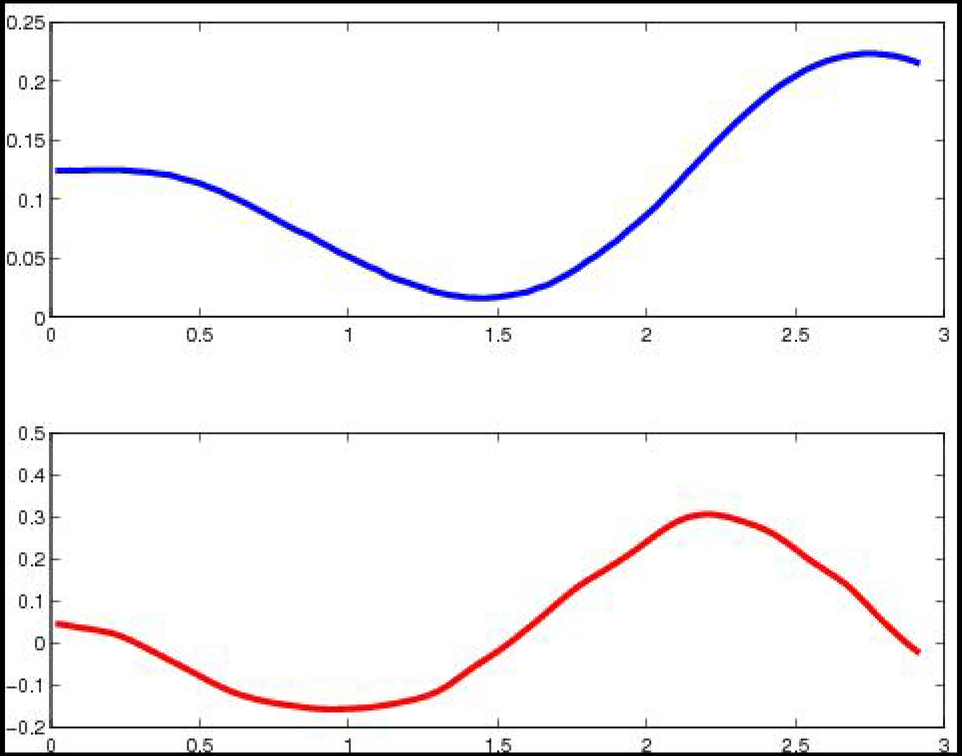
Leading edge flap



Outputs

Angle of attack

Pitch rate

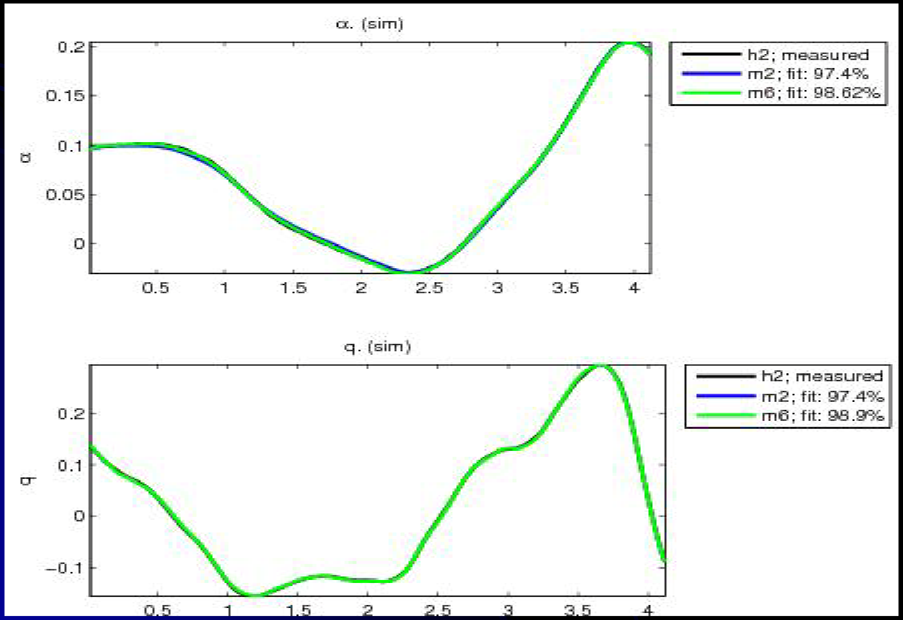


Model and Measured Output

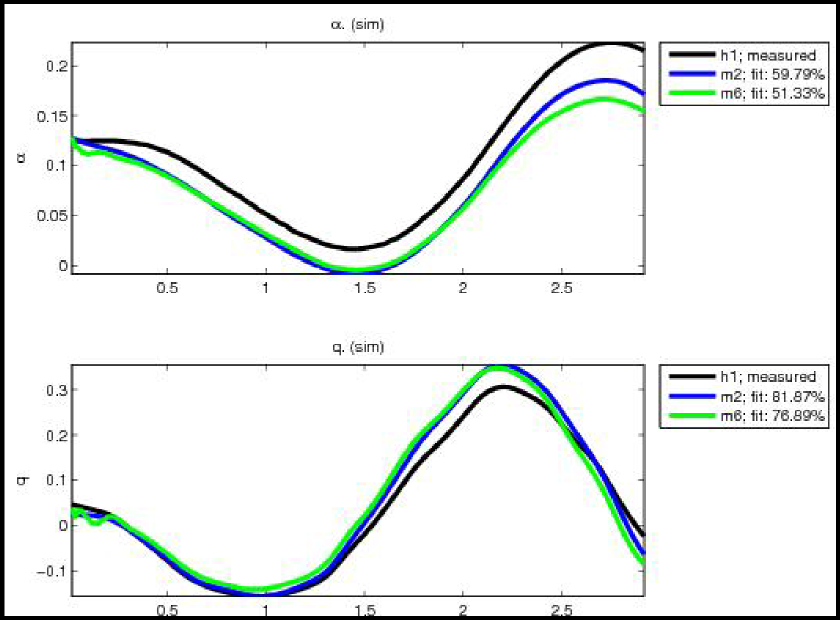
State space models of order 2 and 6:

m2=pem(data,2); m6= pem(data,6)

Estimation data



Validation data



System Identification - Future: Open Areas

- Spend more time with our neighbours!
- Issues in identification of nonlinear systems
- Meet demands from industry
- Convexification

System Identification - Future: Open Areas

- Spend more time with our neighbours!
  - Report from a visit later on
- Issues in identification of nonlinear systems
- Meet demands from industry
- Convexification

Nonlinear Systems

- A user's guide to nonlinear model structures suitable for identification and control:

A "non-elephant zoology" (Ulam)

A Quick Taxonomy of NL Models

1. Black Models:

$$\hat{y}(t|\theta) = \tilde{f}(Z^{t-1},\theta) = f(x(t),\theta)$$

$$x(t) = x(Z^{t-1}) \quad \text{"state" of fixed dimension}$$

$$f(x,\theta) = \sum_{k=1}^{d} \alpha_k g_k(x)$$

$$g_k(x,\theta) = \kappa(\beta_k(x-\gamma_k)), \quad \kappa: \text{ unit function}$$

- The whole ANN, neuro-fuzzy, LS-SVM,etc business

2. Off-white Models: Result from careful modeling from first principles, with certain unknown physical constants being the parameters

Various Shades of Grey …

- 3. Composite Local Models (Local Linear Models)

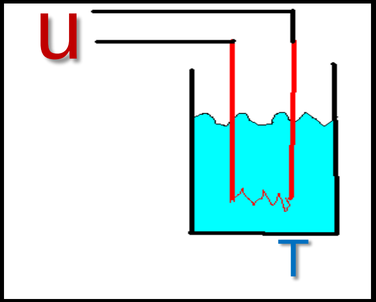
$$\hat{y}(t,\theta,\eta) = \sum_{k=1}^{d} w_k(\rho(t),\eta)\varphi^T(t)\theta^{(k)}$$

- 4. Semi-physical models (nonlinear transformations of measured data based on simple insights)

Semiphysical Modeling

No more than 2 minutes using only highschool physics

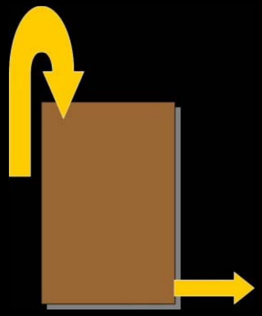
- A simple example



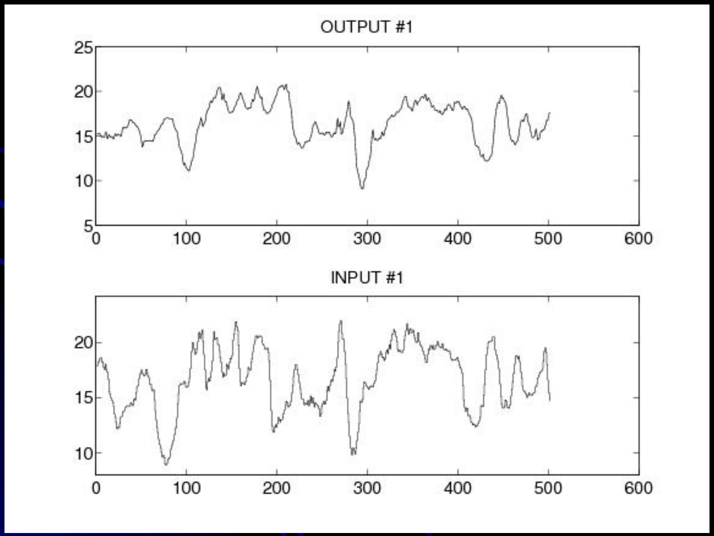
- Input: heater voltage u
- Output: Fluid temperature T

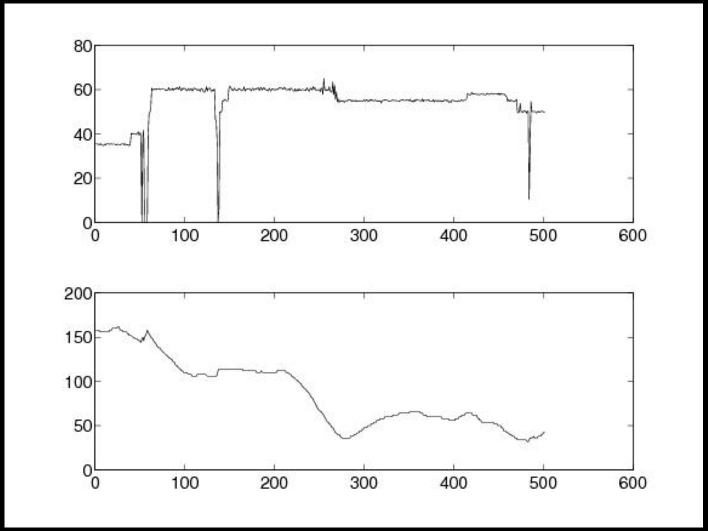
- Square the voltage: $u \rightarrow u^2$

Example: Buffer Vessel Dynamics



- Kappa number of outflow
- Kappa number of inflow
- Flow
- Volume





Model Based on Raw Data

Validation data

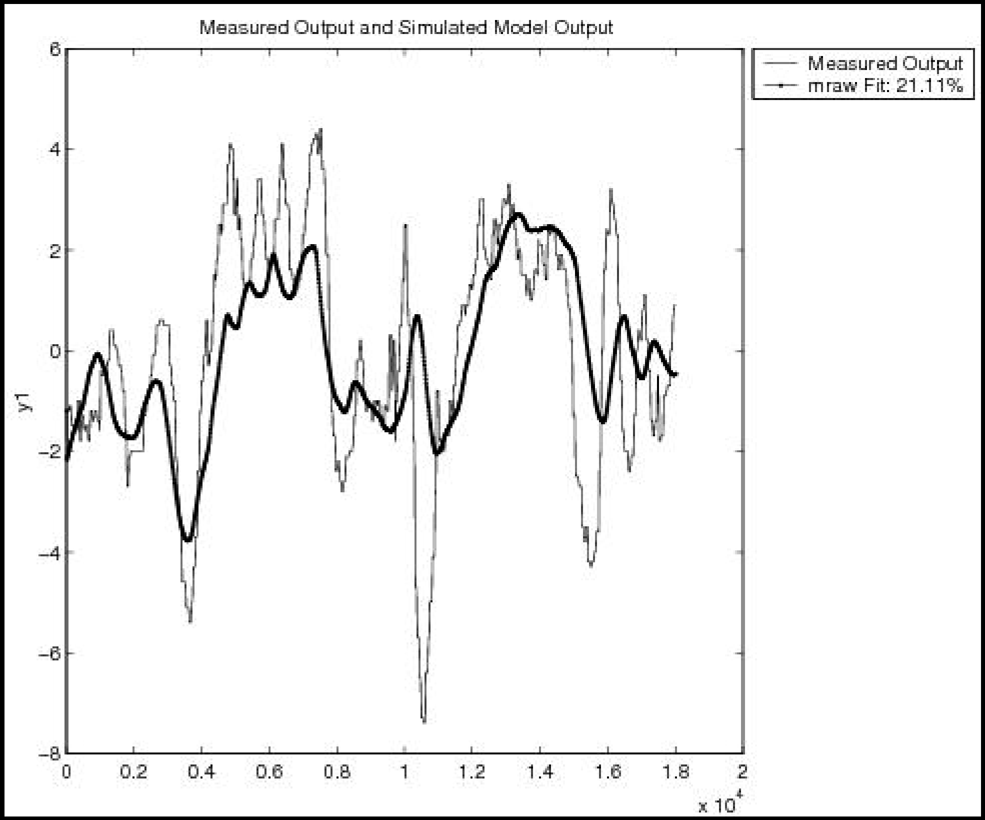
Thin line:

Measured Output

Thick Line:

Simulated Model Output

$$G(s) = \frac{0.818}{1 + 676s} e^{-480s}$$



Now, it's Time to ….

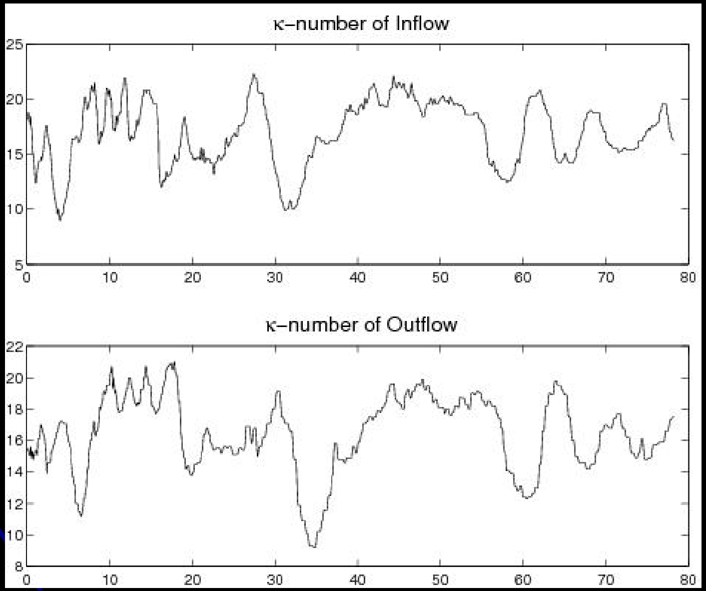
- Think: ….

- If no mixing in tank ("plug flow") a particle that enters the top will exit T seconds later. T = (Tank Volume)/(Flow)

$$\left[\frac{m^3}{m^3/s} = s\right]$$

Resample Data!

```
z = [y,u]; pf = flow./level;
t = 1:length(z)
newt = interp1([cumsum(pf),t],[pf(1):sum(pf)]');
newz = interp1([t,z], newt);
```



Semi-physical Model with resampled data:

Validation data:

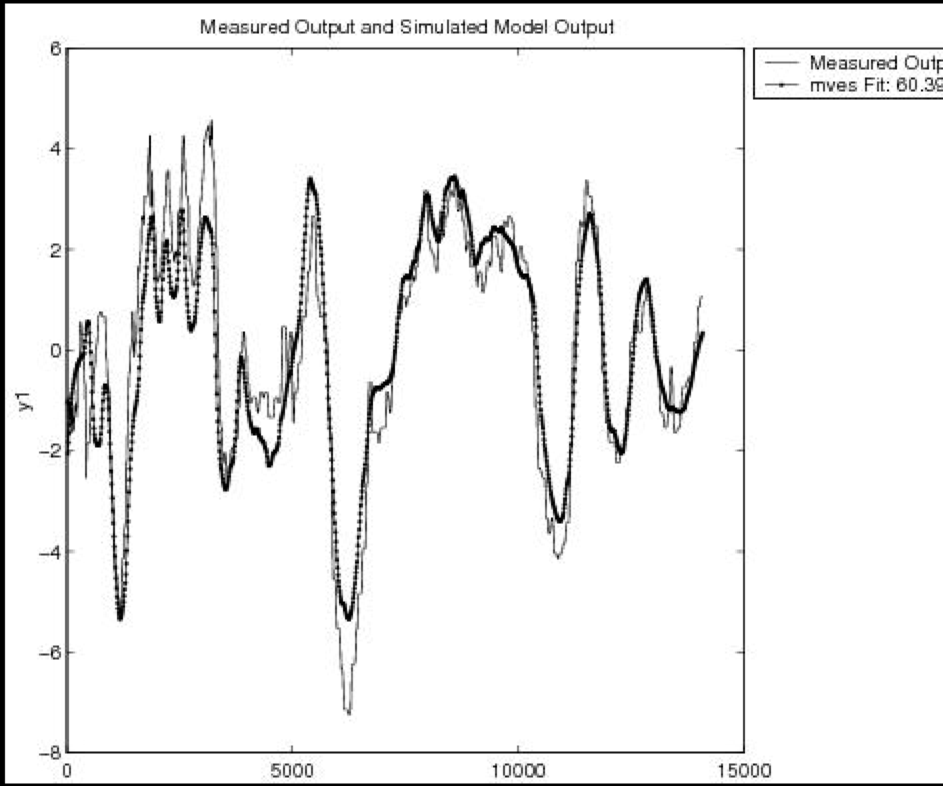
Thin line:

Measured Output

Thick Line:

Simulated Model Output

$$G(s) = \frac{0.8116}{1 + 110.28s} e^{-369.58s}$$



System Identification - Future: Open Areas

- Spend more time with our neighbours!
- Issues in identification of nonlinear systems
- Meet demands from industry
- Convexification

Industrial Demands

- Data mining in large historical process data bases ("K,M,G,T,P")

  All process variables, sampled at 1 Hz for 100 years
  = 0.2 PByte

PM 12, Stora Enso Borlänge
75000 control signals, 15000 control loops

- A serious integration of physical modeling and identification (not just parameter optimization in simulation software)

Industrial Demands: Simple Models

- Simple Models/Experiments for certain aspects of complex systems
- Use input that enhances the aspects, …
- … and also conceals irrelevant features
  - Steady state gain for arbitrary systems
    - Use constant input!
  - Nyquist curve at phase crossover
    - Use relay feedback experiments
  - But more can be done …

System Identification - Future: Open Areas

- Spend more time with our neighbours!
  - Report from a visit later on
- Issues in identification of nonlinear systems
- Meet demands from industry
- Convexification
  - Formulate the estimation task as a convex optimization problem

Convexification I

Example:

Michaelis – Menten kinetics

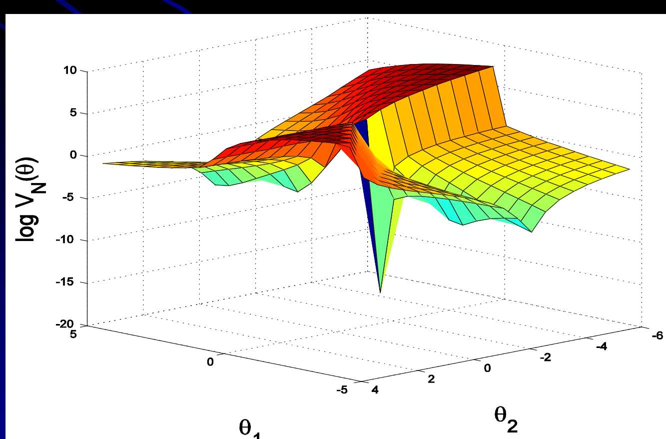
Are Local Minima an inherent feature of a model structure?

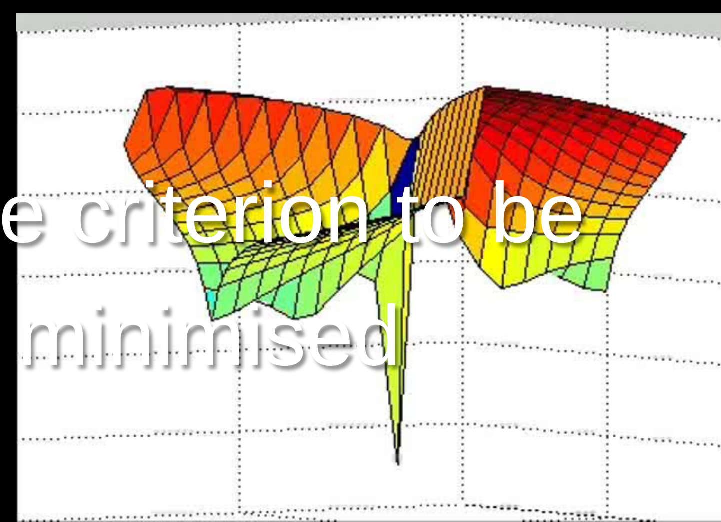
$$\dot{y} = \theta_1 \frac{y}{\theta_2 + y} - y + u$$

$$y_m(t_k) = y(t_k) + e(t_k)$$

$$\dot{\hat{y}}(t|\theta) = \theta_1 \frac{\hat{y}(t|\theta)}{\theta_2 + \hat{y}(t|\theta)} - \hat{y}(t|\theta) + u(t)$$

$$V_N(\theta) = \sum_{k=1}^{N} (y_m(t_k) - \hat{y}(t_k|\theta))^2$$





The criterion to be minimised

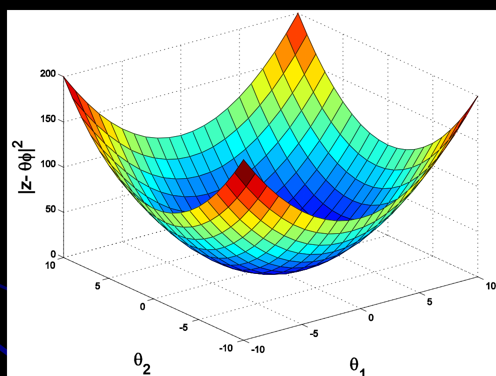
Massage the equations:

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

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

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

$$z = \theta \phi$$



This equation is a linear regression that relates the unknown parameters and measured variables. We can thus find them by a simple least squares procedure. We have, in a sense, convexified the problem

Is this a general property?

Yes, any identifiable structure can be rearranged as a linear regression (Ritt's algorithm)

Convexification II
Manifold Learning

$$\mathcal{X} \rightarrow g(x) \rightarrow \mathcal{Z} \rightarrow h(z) \rightarrow \mathcal{Y}$$

X : Original regressors

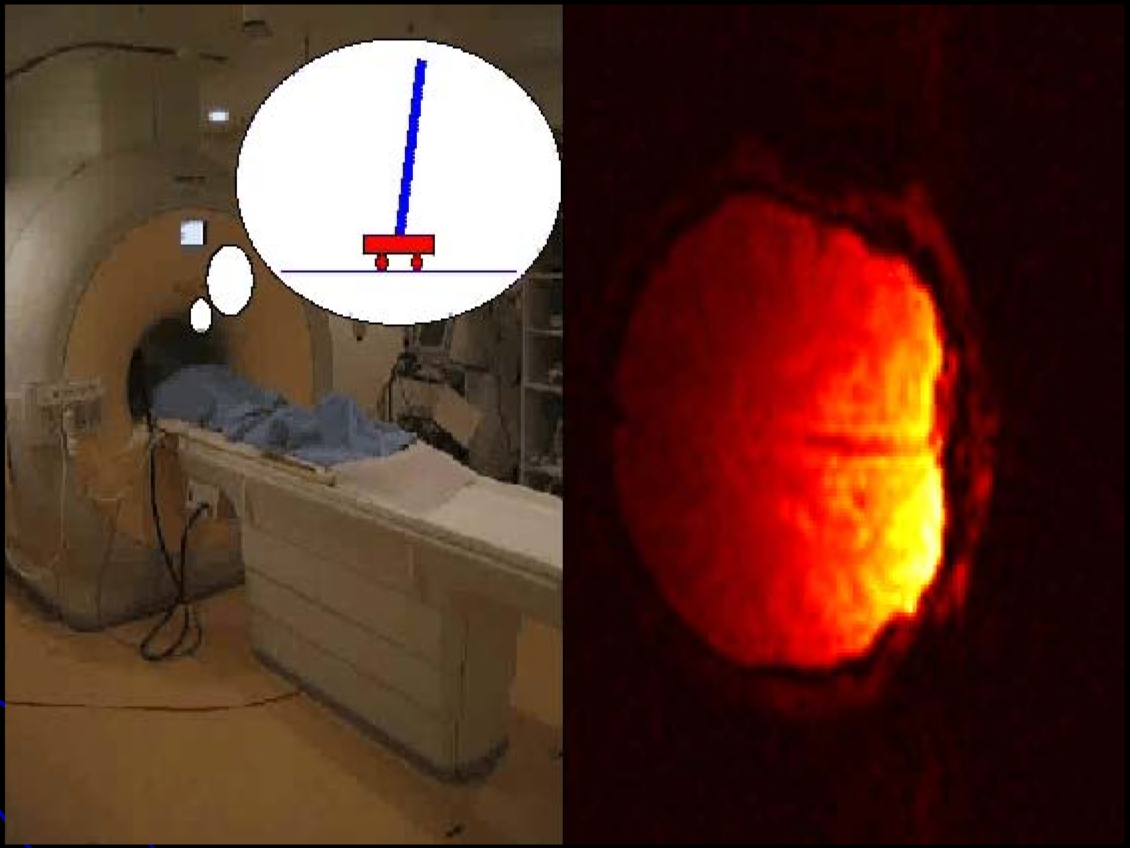
g(x) Nonlinear, nonparametric recoordinatization

Z : New regressor, possibly of lower dimension

h(z): Simple convex map

Y : Goal variable (output)

Analysis of fMRI signals



The observed data

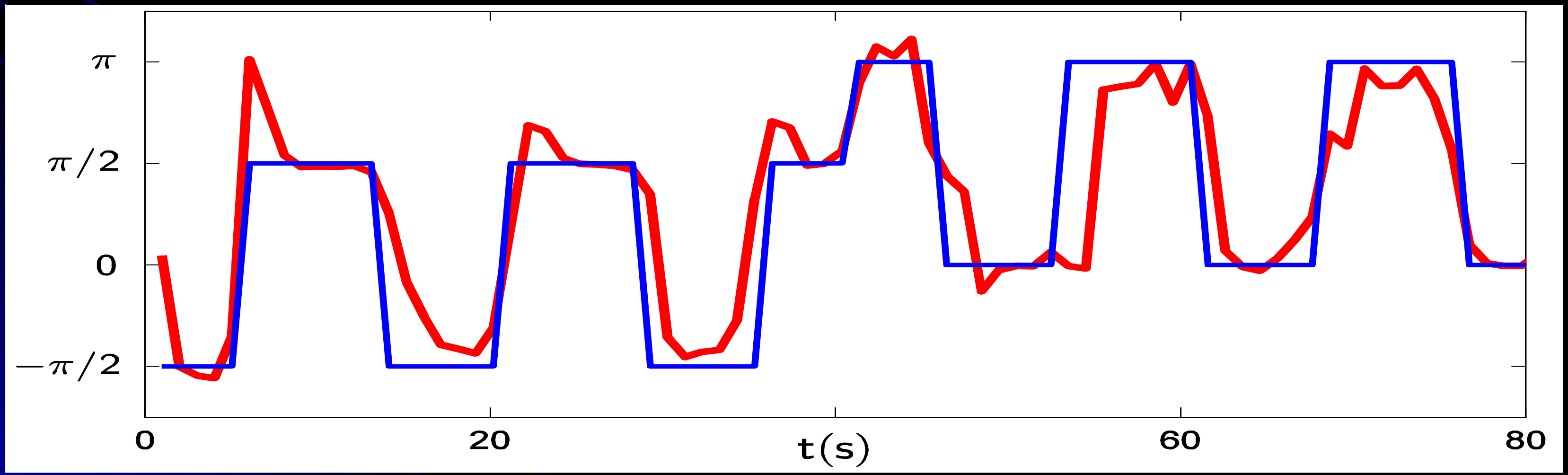
The patient in the magnet camera is moving his eye focus in a circle left - right - up - down. 128 voxels in the visual cortex are monitored by fMRI, giving a vector $\varphi(t) \in R^{128}$ *sampled every* two seconds. The output y(t) is the viewing angle $y(t) \in [-\pi, \pi]$

The regressor $\varphi(t)$ *is 128-dimensional. At the same time the "brain* activity is 1-dimensional", so the interesting variation in the regressor space should be confined to a one-dimensional manifold

WDMR: Estimated model

We have devised a method, WDMR, that is based on LLE (Local Linear Embedding) for estimating a low dimensional manifold, and finds a function from this manifold to the observed outputs.

Below we show the predicted y-values (angles $[-\pi, \pi]$) *(red) for* validation measurements together with the corresponding true angles (blue).



Conclusions

- System identification is a mature subject ...
  - 50 years old, many publications and the longest running symposium series
- … and much progress has allowed important industrial applications …
- … but it still has an exciting and bright future!

Thanks

Research: Martin Enqvist, Torkel Glad, Håkan Hjalmarsson, Henrik Ohlsson, Jacob Roll

Discussions: Bart de Moor, Johan Schoukens, Rik Pintelon, Paul van den Hof

Comments on presentation: Martin Enqvist, Håkan Hjalmarsson, Kalle Johansson, Ulla Salaneck, Thomas Schön, Ann-Kristin Ljung

Special effects: Effektfabriken AB, Sciss AB