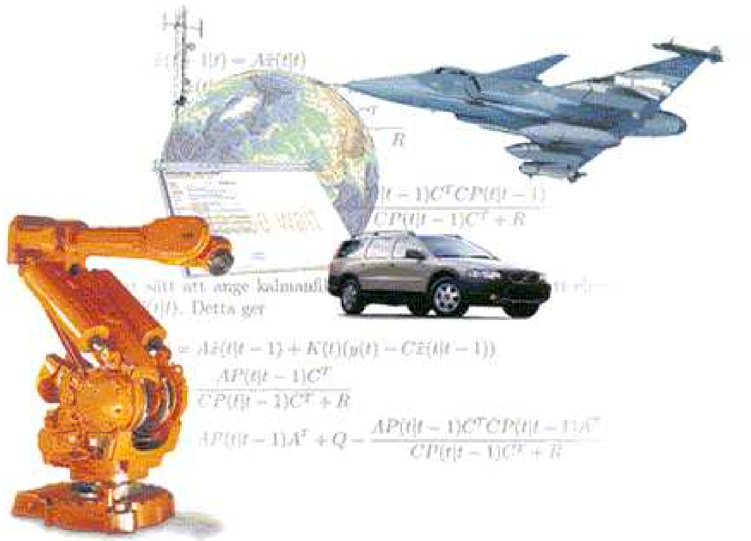
System Identification
From Data to Models

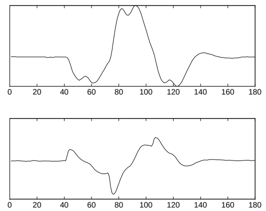


Lennart Ljung

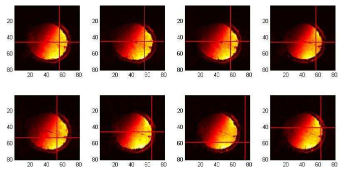
Division of Automatic Control
Linköping University
Sweden

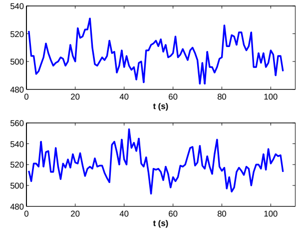
The Problem

Aircraft Dynamics:

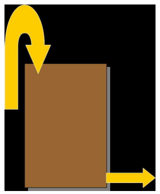


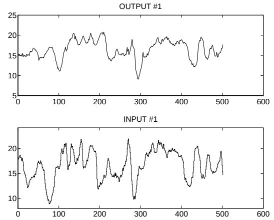
Brain Activity (fMRI):



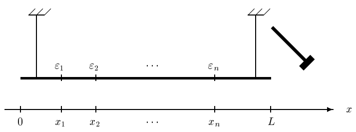


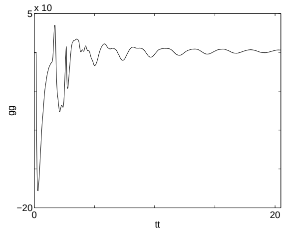
Pulp Buffer Vessel:





Viscoelasticity:





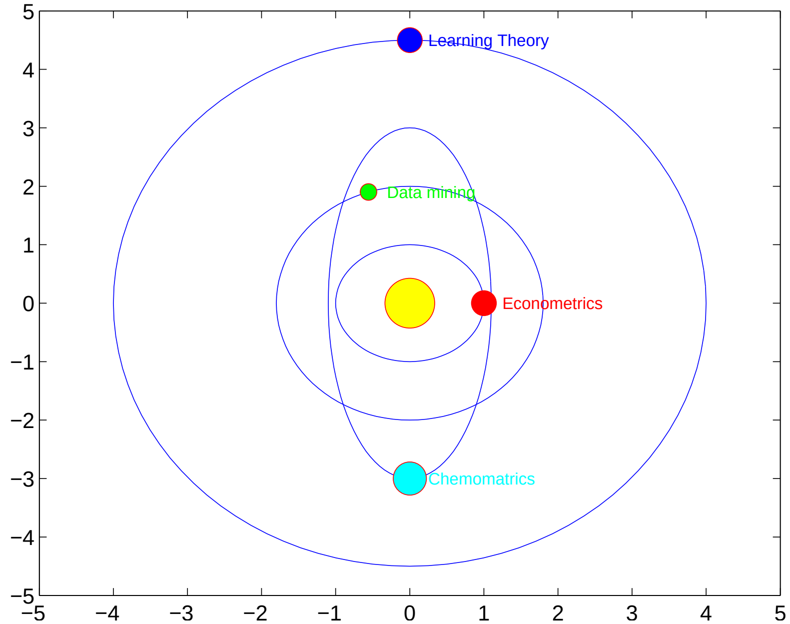
The Confusion

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 * Cramér - Rao * Identification for Control * N4SID * Experiment Design * Fisher Information * Local Linear Models * Kullback-Liebler Distance * Maximum Entropy * Subspace Methods * Kriging * Gaussian Processes * Ho-Kalman * Self Organizing maps * Quinlan's algorithm * Local Polynomial Models * Direct Weight Optimization * 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 * Mutivariate Analysis * Missing Data * Density Estimation * PEM *

The Communities

Constructing (mathematical) 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.



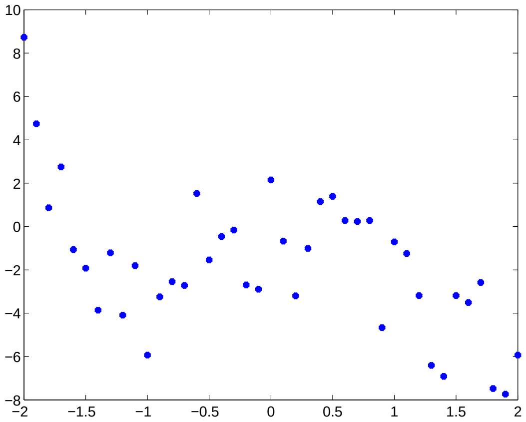
The Core

Central terms

- Model $\mathfrak{m}$ – Model Class $\mathcal{M}$ – Complexity (Flexibility) $\mathcal{C}$
- Information $\mathcal{I}$ – Data Z
- Estimation – Validation (Learning – Generalization)
- Model fit $\mathcal{F}(\mathfrak{m}, Z)$

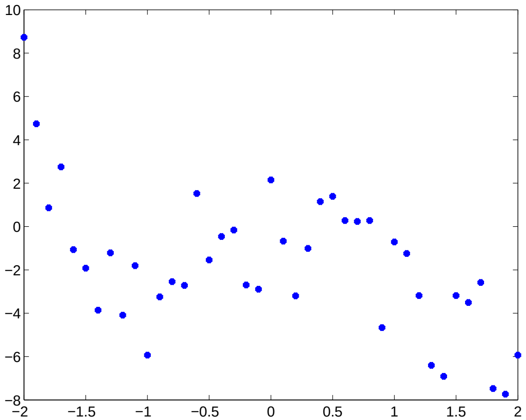
Estimation

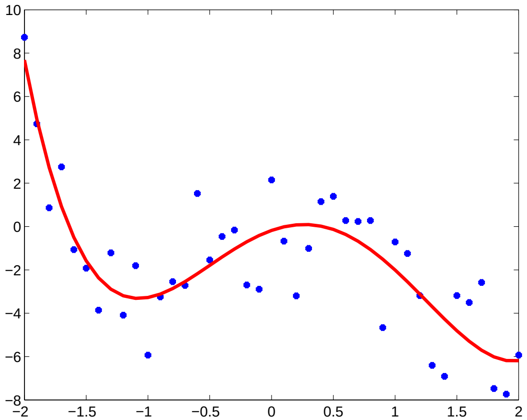
information in data



Estimation

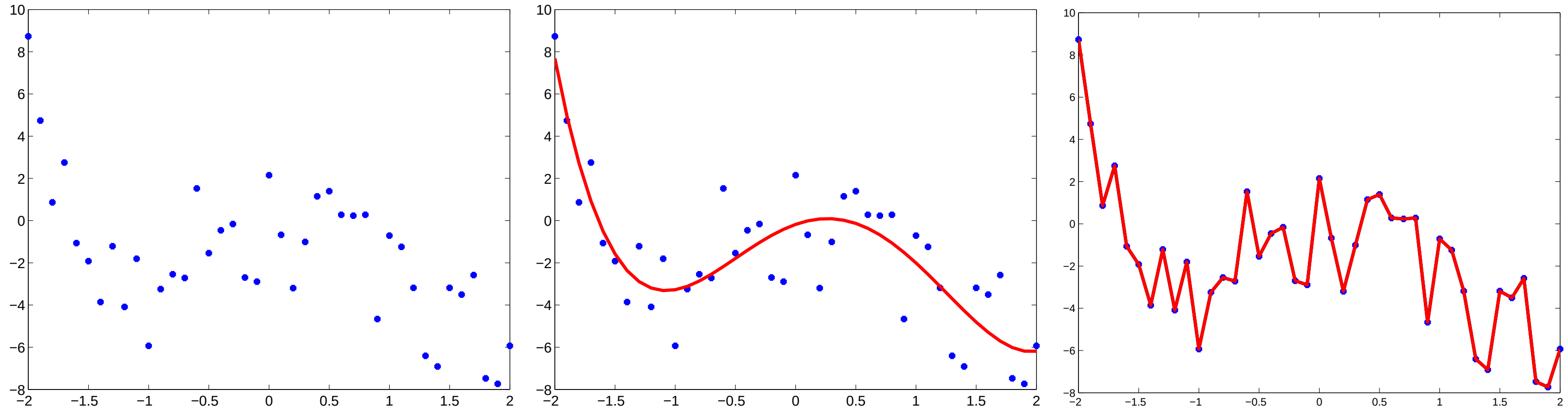
Squeeze out the relevant information in data





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!

Nature is Simple! (Occam's razor, Lex Parsimoniae...)

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

So, conceptually:

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

Examples:

- Search for a model in sets with a maximal Complexity
- (Akaike):
  $\hat{\mathfrak{m}} = \arg\min \log[\sum \varepsilon^2(t,\theta)] + 2 \dim \theta$ ε: Model error θ:Model parameters
- (Regularization): $\hat{\mathfrak{m}} = \arg\min \sum \varepsilon^2(t,\theta) + \delta \|\theta\|^2$

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 data in a flexible model set!
(Elephant #1)

Bias and Variance

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

E: Expected Value

Then

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

MSE $=$ B: BIAS $+$ V: Variance

Error: $=$ Systematic $+$ Random

$\hat{\mathfrak{m}} \in \mathcal{M}$: As $\mathcal{C}(\mathcal{M})$ increases, B decreases $\&$V increases

This bias/variance trade-off is at the heart of estimation.

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

The value of information in data depends on prior knowledge.
Observe Y. Let its probability density function (pdf) 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
  - ....
  - Bootstrap
  - Regularization to control complexity (LASSO, LARS,...)
- Econometrics
  - Volatility Clustering (varying variance)
  - Common roots for variations (co-integration)
- Statistical Learning Theory
  - Convex Formulations, SVM
  - VC-dimensions
- Machine Learning
  - Self-organizing maps, logical trees
  - Grown out of artificial intelligence, more and more statistically oriented

The Communities Around the Core II

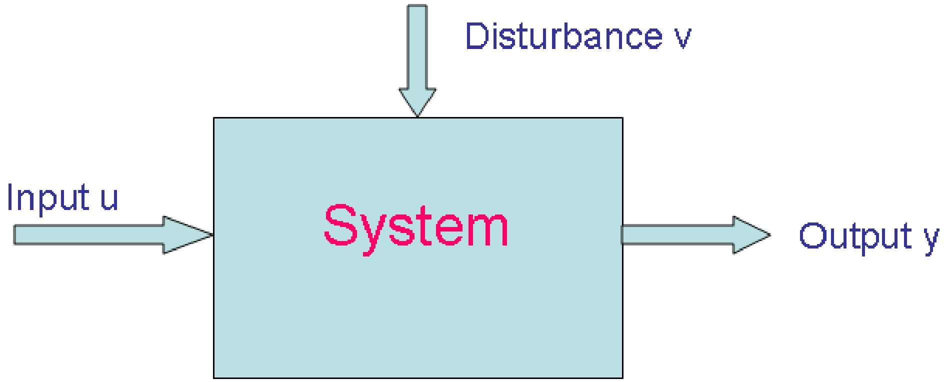
- Manifold Learning
  - Observed data belongs to a high-dimensional space
  - The action takes place on a lower dimensional manifold: Find that!
- Chemometrics – Statistical Process Control
  - High-dimensional Data Spaces (Many process variables)
  - Find linear low dimensional subspaces that capture the essential state
  - PCA, PLS (Partial Least Squares), ...
- Data Mining
- Artificial Neural Networks
- ....
- ...

- Another satellite encircling the core.
- Deals with mathematical models of dynamic systems.
- Term used in the automatic control community (coined by Lotfi Zadeh 1956)
- Typical themes:
  - Useful model structures
  - Adapt and adopt the core's fundamentals
  - Experiment design (make $\mathcal{I}$ large)
    - with intended model use in mind ("identification for control")

A Dynamic system has an output response y that depends on (all) previous values of an input signal u. It is also typically affected by a disturbance signal v. So the output at time t can be written as

$$y(t) = g(u^t, v^t)$$

where superscript denotes the signal's values from the remote past up to the indicated time. The input signal u is known (measured), while the disturbance v is unmeasured.



More Formalized Questions

Think discrete time data sequences:
$Z^t = [u(1), u(2), ..., u(t), y(1), y(2), ..., y(t)]$
We need to get hold of a “simulation function”

$$y(t) = g(u^t)$$

and/or a prediction function

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

Note that $\tilde{f} \Rightarrow g$

The predictor function $\hat{y}(t|t-1) = \tilde{f}(Z^{t-1})$ is what we try to estimate from data. It is (partly) unknown, so parameterize it within a certain model class $\mathcal{M}$:

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

Generic way to estimate θ:

$$\hat{\theta}_N = \arg\min_{\theta} \sum_{t=1}^{N} \|y(t) - \hat{y}(t|\theta)\|^2$$

Two main model classes:

- Linear: $\tilde{f}$ linear in Z:
- Nonlinear: $\tilde{f}$ nonlinear in Z

Linear Dynamic Models

$$\tilde{f}(Z^{t-1}, \theta) = \hat{y}(t|\theta) = \sum_{k=1}^{\infty} \tilde{g}_k(\theta)u(t-k) + \sum_{k=1}^{\infty} \tilde{h}_k(\theta)y(t-k)$$

$$= \tilde{G}(q, \theta)u(t) + \tilde{H}(q, \theta)y(t)$$

$$qy(t) = y(t+1) \quad \text{(shift operator)}$$

This is how the output is predicted. Equivalent to assume that y is generated from

$$y(t) = G(q, \theta)u(t) + H(q, \theta)e(t) \quad \text{where } e \text{ is white noise}$$

$$\tilde{G} = H^{-1}G \quad \tilde{H} = I - H^{-1}$$

Parameterization of Linear Dynamic Models

So linear dynamic models can be written in transfer function form

$y(t) = G(q,\theta)u(t) + H(q,\theta)e(t)$ G and H functions of the delay operator q

Typical parameterizations: rational functions in q (Black-Box)

$$G(q,\theta) = \frac{b_1 q^{-1} + \ldots + b_n q^{-n}}{1 + a_1 q^{-1} + \ldots + a_n q^{-n}}$$

State Space (Grey-Box, originating from a system of first order ODEs)

$$x(t+1) = A(\theta)x(t) + B(\theta)u(t)$$
$$y(t) = C(\theta)x(t)$$

$$G(q,\theta) = C(\theta)(qI - A(\theta))^{-1}B(\theta)$$

Five inputs and two outputs.

Build models of the kind

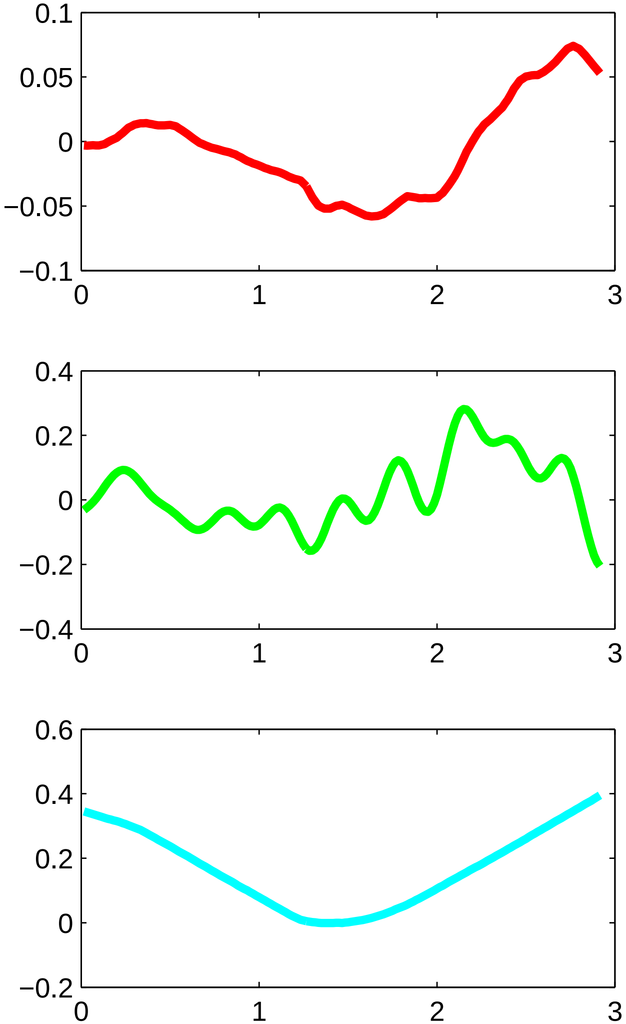
$$x(t+1) = Ax(t) + Bu(t) + Ke(t)$$

$$y(t) = Cx(t) + e(t)$$

“order” = dim x.

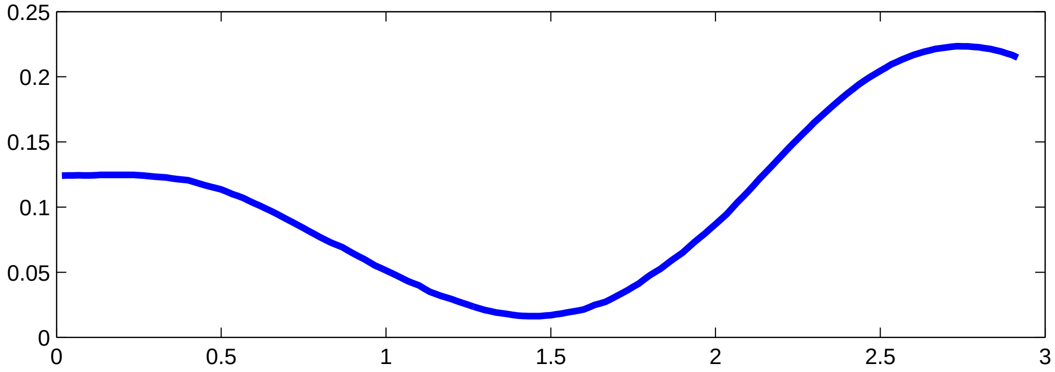
Inputs

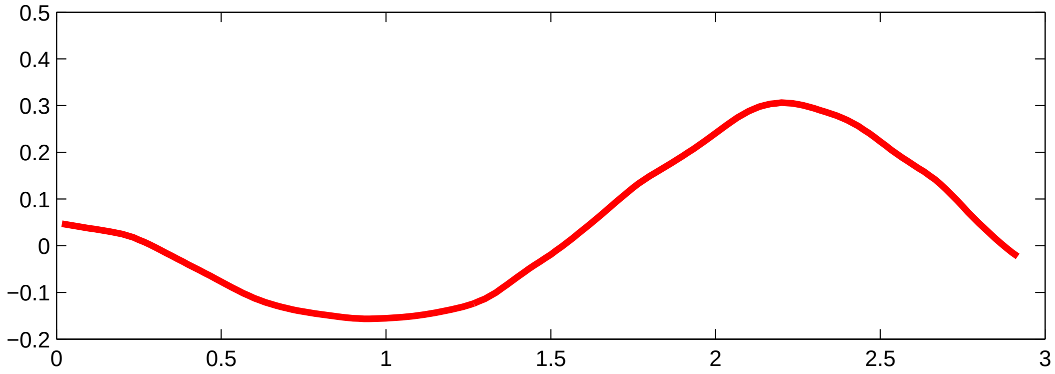
Elevator, $\frac{d}{dt}$elevator, leading edge, canard, $\frac{d}{dt}$canard



Outputs

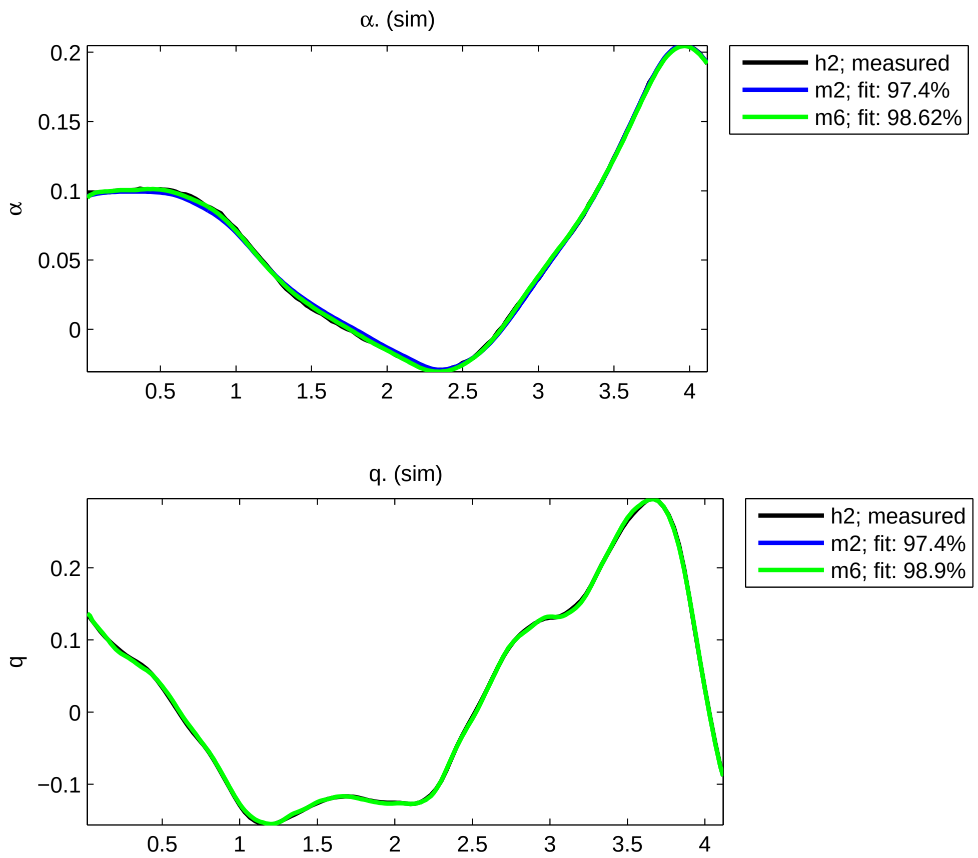
Angle of attack and Pitch rate





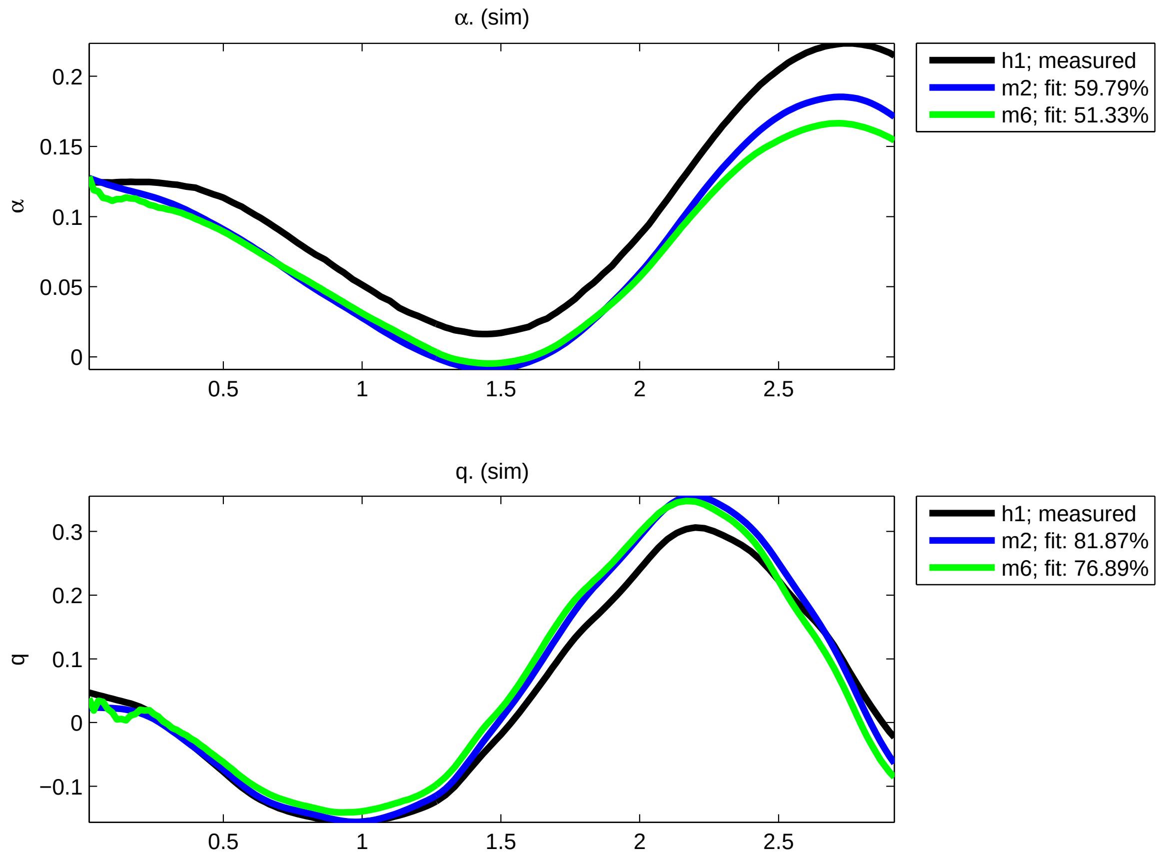
State space models of orders 2 and 6:

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



Linear Models: Fit to validation data

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



A Quick Classification of Non-Linear Models: B/W

"A non-elephant zoology" (Ulam)

1. Black: Basis-function expansion models

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

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

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

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

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

2. Off-white: result from careful physical modelling from first principles, with certain unknown physical constants being the parameters.

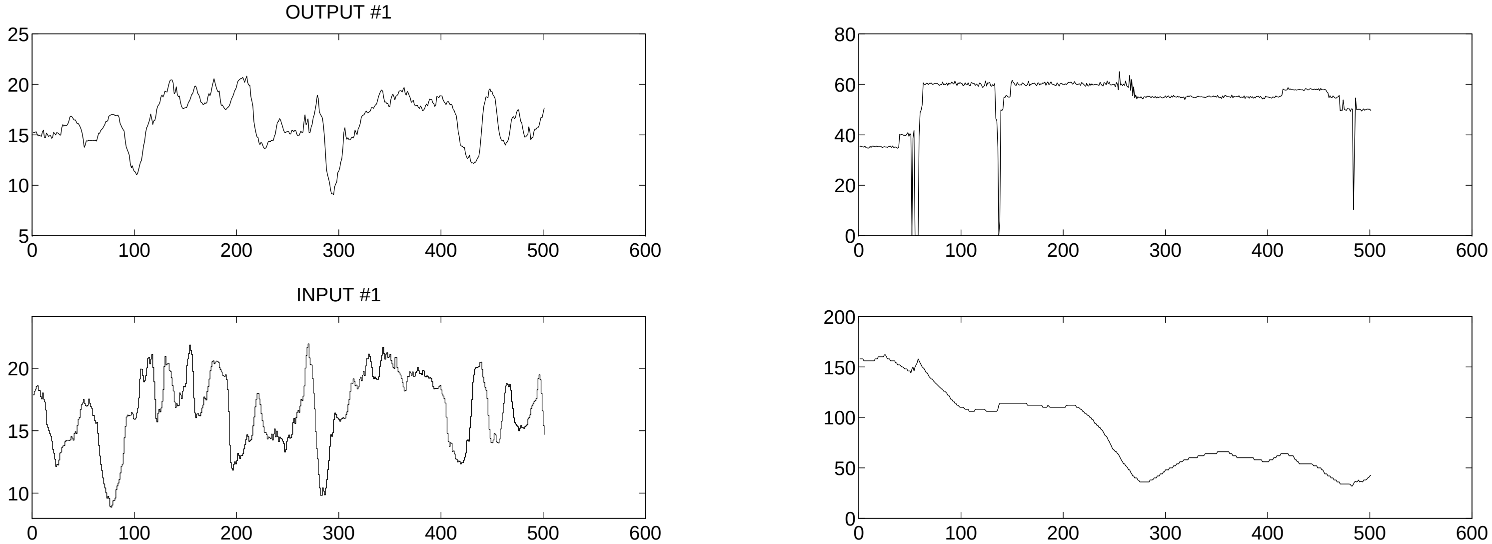
3. Composite Local models (local linear models)

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

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

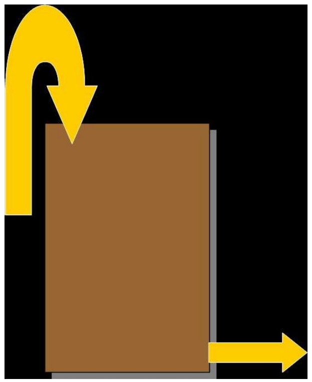
- Probably the most common nonlinear models in industrial practice

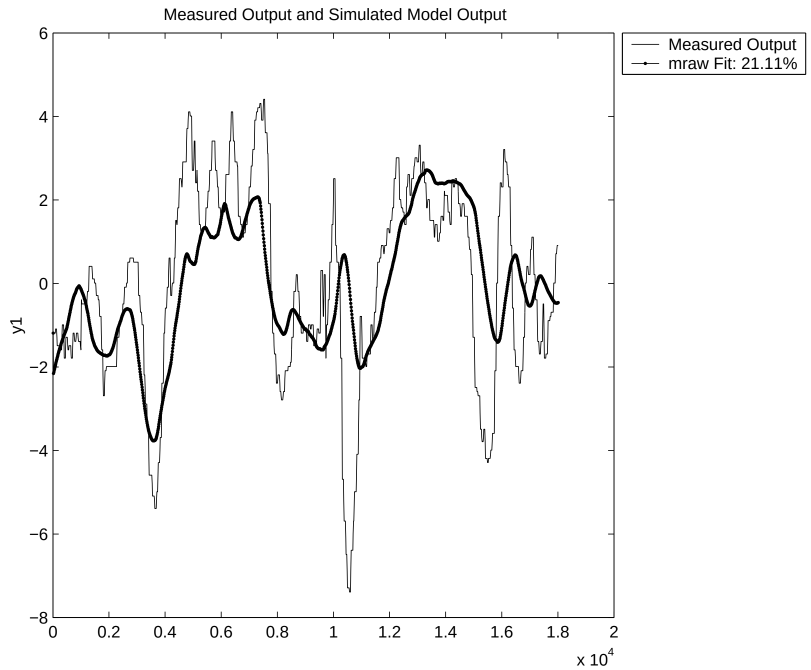
Buffer Vessel Dynamics



κ-number of outflow,
κ-number of inflow,

flow
volume





Dashed line: κ-number after the vessel, actual measurements.
Solid line: Simulated κ-number using the input only and a process model estimated using the first 200 data points. $G(s) = \frac{0.818}{1+676s} e^{-480s}$

Now it's time to

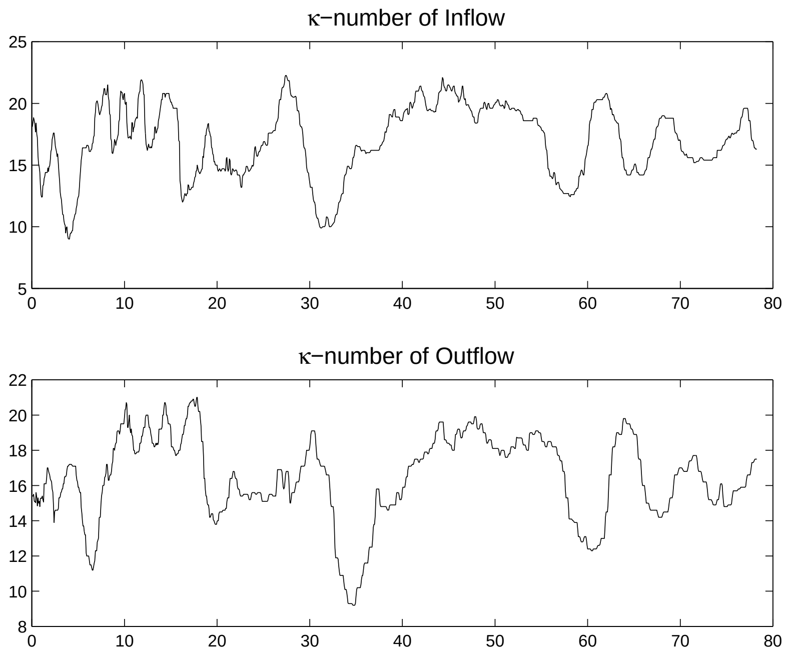
Think: ....

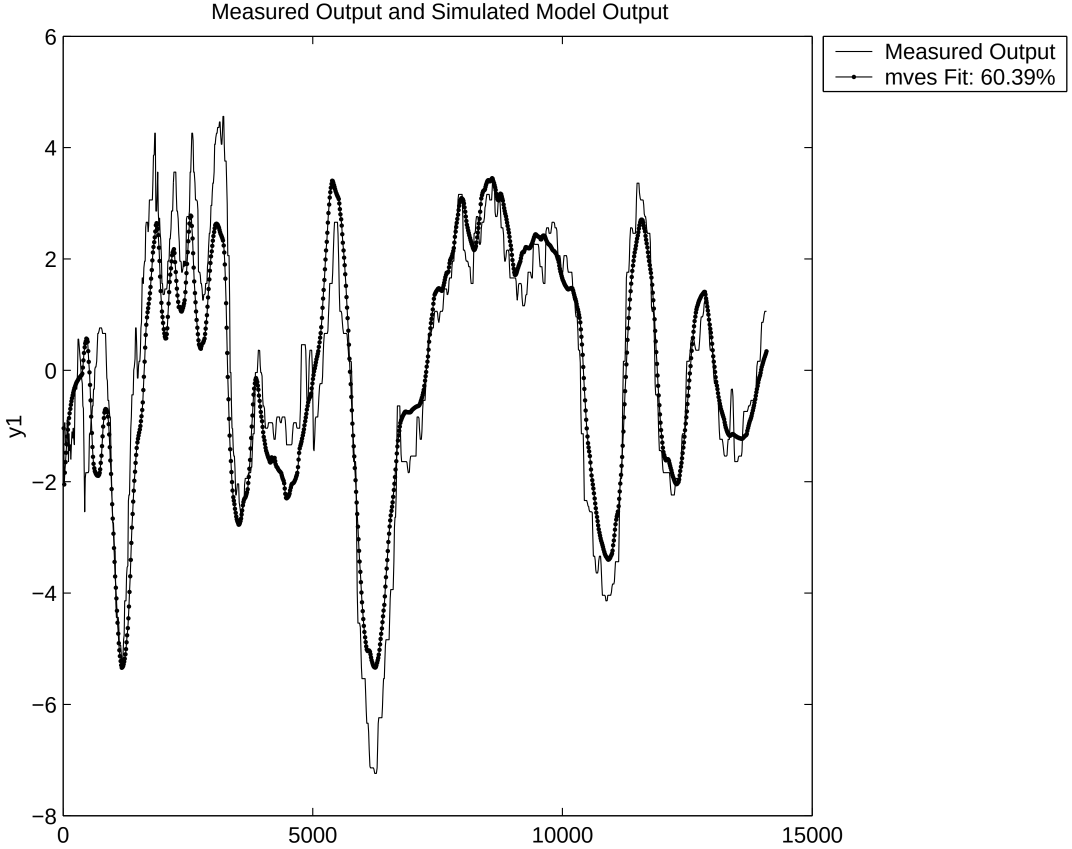
If no mixing in tank ("plug flow") a particle that enters the top will exit T seconds later, where

$$T = \frac{\text{Tank Volume}}{\text{Flow}} : \left[\frac{m^3}{m^3/s} = s\right]$$

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);
```





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

Conclusions

- Very rich literature on building models from data (“The communities”)
- Relatively few leading principles (“The core”)
- System Identification deals with building models of dynamic systems
  - Parameterization of linear and nonlinear dynamic models
  - ... with and without physical insight
  - ... and associated algorithms
  - Influence of experiment design for model quality