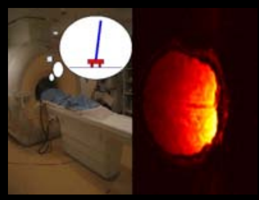
Perspectives on System Identification

Lennart Ljung
Linköping University, Sweden

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 Identfication.

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

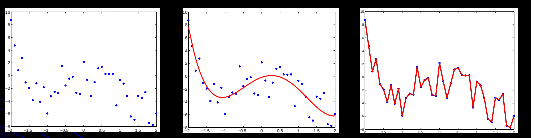
Estimation – Validation (Learning – Generalization)

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:

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

- Ex: Akaike: $\hat{\mathfrak{m}} = \arg\min_{\mathfrak{m}\in\mathcal{M}} \log \sum \varepsilon^2(t,\theta) + 2\dim\theta$

θ : model parameters, ε : model error

- Regularization: $\hat{\mathfrak{m}} = \arg\min_{\mathfrak{m}\in\mathcal{M}} \sum \varepsilon^2(t,\theta) + \delta\|\theta\|^2$

Estimation and Validation

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

$F(\hat{\mathfrak{m}}, Z_e^N)$ ("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!

Bias and Variance

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

$\hat{\mathfrak{m}} \in \mathcal{M}$: 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

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

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 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 Proc., HMM, 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")

System Identification - Future: Open Areas

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

Model Reduction

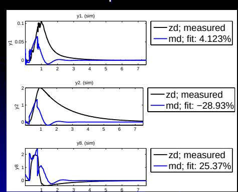
System Identification is really "System Approximation" and therefore closely related to Model Reduction.

Model Reduction is a separate area with an extensive literature (``another satellite''), which can be more seriously linked to the system identification field.

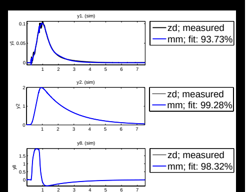
- **Linear systems - linear models**
  - Divide, conquer and reunite (outputs)!
- **Non-linear systems – linear models**
  - Understand the linear approximation - is it good for control?
- **Nonlinear systems -- nonlinear reduced models**
  - Much work remains

Linear Systems - Linear Models
Divide – Conquer – Reunite!

Helicopter data: 1 pulse input; 8 outputs (only 3 shown here).
State Space model of order 20 wanted.
First fit all 8 outputs at the same time:



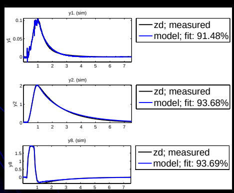
Next fit 8 SISO models of order 12, one for each output:



Linear Systems - Linear Models
Divide – Conquer – Reunite!

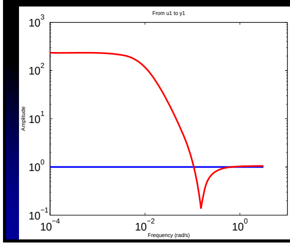
Now, concatenate the 8 SISO models, reduce the 96th order model to order 20, and run some more iterations.

(mm = [m1;…;m8]; mr = balred(mm,20); model = pem(zd,mr); compare(zd,model))



Linear Models from Nonlinear Systems

System: $y(t) = u(t) + 0.01u^3(t)$,
u non-Gaussian $|u(t)| \leq 3$ (Martin Enqvist)
Model : $y(t) = G(q,\theta)u(t) + e(t)$: `m = oe(z,[2 2 1])`



Red: Amplitude Bode plot for estimated model
Blue: Model without $0.01u^3(t)$
Output discrepancy $\leq 1\%$

Model reduction from nonlinear to linear could be surprising!

Nonlinear Systems

- A user's guide to nonlinear model structures suitable for identification and control
- Unstable nonlinear systems, stabilized by unknown regulator

- Stability handle on NL blackbox models

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
  = 0.1 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

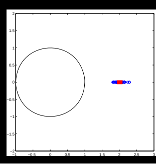
- Simple Models/Experiments for certain aspects of complex systems
- Use input that enhances the aspects, …
- … and also conceals irrelevant features
  - Steady state gain for arbitrary systems
    - Use constant input!
  - Nyquist curve at phase crossover
    - Use relay feedback experiments
  - But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

An Example of a Specific Aspect

- Estimate a non-minimum-phase zero in complex systems (without estimating the whole system) – For control limitations.
- A NMP zero at α for an arbitrary system can be estimated by using the input

$$u = \frac{c}{z^{-1} + \alpha} e$$

Example: 100 complex systems, all with a zero at 2, are estimated as 2nd order FIR models $y(t) = b_1 u(t) + b_2 u(t-1)$



System Identification - Future: Open Areas

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

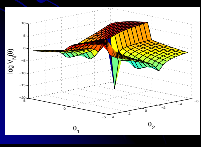
Convexification I

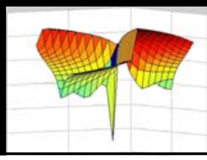
Example:
Michaelis – Menten kinetics

Are Local Minima an Inherent feature of a model structure?

$$\dot{y} = \theta_1 \frac{y}{\theta_2 + y} - y + u$$
$$y_m(t_k) = y(t_k) + e(t_k)$$

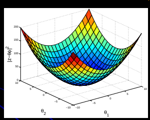
$$V_N(\theta) = \sum_{k=1}^{N} (y_m(t_k) - \hat{y}(t_k|\theta))^2$$
$$\dot{\hat{y}}(t|\theta) = \theta_1 \frac{\hat{y}(t|\theta)}{\theta_2 + \hat{y}(t|\theta)} - \hat{y}(t|\theta) + u(t)$$





Massage the equations:

$$\dot{y} = \theta_1 \frac{y}{\theta_2 + y} - y + u$$
$$\dot{y}y + \theta_2 \dot{y} = \theta_1 y - y^2 - \theta_2 y + uy + \theta_2 u$$
$$\text{or } \dot{y}y + y^2 - uy = [\theta_1 \quad \theta_2] \begin{bmatrix} y \\ u - \dot{y} - y \end{bmatrix}$$
$$z = \theta\phi$$



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

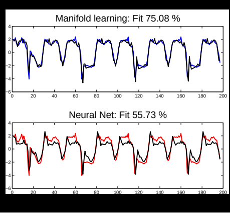
h(z): Simple convex map

Y: Goal variable (output)

Narendra-Li's Example

$$x_1(t+1) = \left(\frac{x_1(t)}{1 + x_1^2(t)} + 1\right) \sin(x_2(t))$$
$$x_2(t+1) = x_2(t)\cos(x_2(t)) + x_1(t)\exp\left(-\frac{x_1^2(t) + x_2^2(t)}{8}\right) + \frac{u^3(t)}{1 + u^2(t) + 0.5\cos(x_1(t) + x_2(t))}$$
$$y(t) = \frac{x_1(t)}{1 + 0.5\sin(x_2(t))}$$

Use LLE for the nonlinear recoordinatization of regressors and use simple linear map. (All convex problems.) Compare with standard ANN (Henrik Ohlsson)



Simulate estimation and validation data $u \rightarrow y$. Define regressors as delayed inputs and output $y(t-k), u(t-k)$ and build a NL ARX model
$y(t) = f(y(t-k), u(t-k), k = 1, \ldots, n)$

Conclusions

- System identification is a mature subject ...
  - same age as IFAC, with the longest running symposium series
- … and much progress has allowed important industrial applications …
- … but it still has an exciting and bright future!