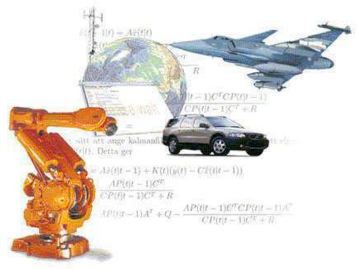
Sysid Course VT1 2016
An Overview.



Lennart Ljung

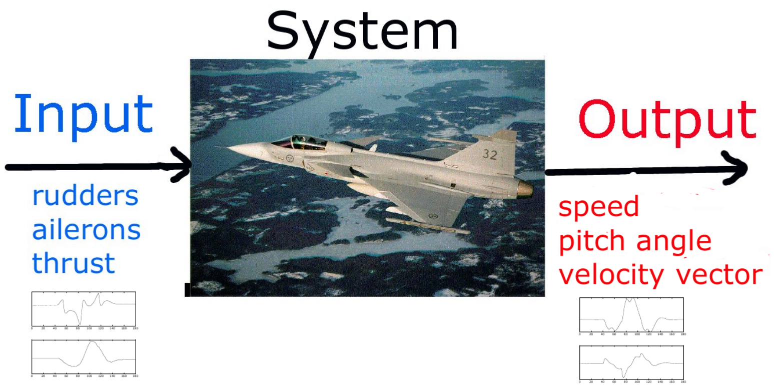
Automatic Control, ISY, Linköpings Universitet

Outline

- An Umbrella Contribution for the Material in the Course
- The classic, conventional System Identification Setup
- The Identification Loop
- Model and Model Structures
- Identification Methods
- Model Validation
- $\boxed{\rightarrow \text{xyz}}$ means: details are given on page xyz in the textbook

  Ljung: System Identification. Theory for the User, 2nd Edition, Prentice-Hall 1999.

An Introductory Example $\boxed{\rightarrow 2{,}530}$



System

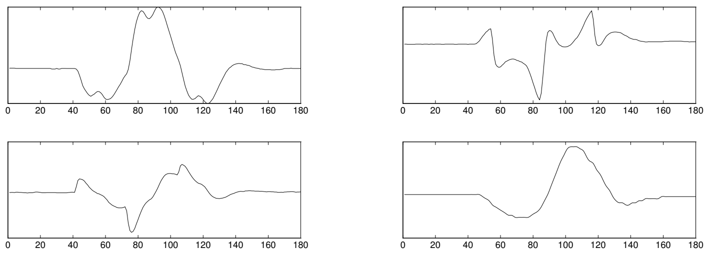
Input

rudders
ailerons
thrust

Output

speed
pitch angle
velocity vector

Data from Gripen



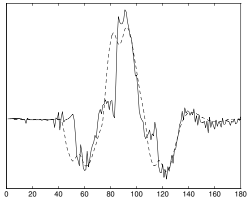
Pitch rate, Canard,
Elevator, Leading Edge Flap

- How do the control surface angles affect the pitch rate?
- Aerodynamical derivatives?
- How to use the information in flight data?

Aircraft Dynamics: From input 1

$y(t)$ pitch rate at time t. $u_1(t)$ canard angle at time t. $T = 1/60$.
Try

$$y(t) = \\ +b_1 u_1(t-T) + b_2 u_1(t-2T) + b_3 u_1(t-3T) + b_4 u_1(t-4T)$$

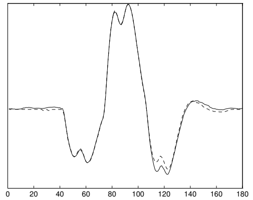


Dashed line: Actual Pitch rate. Solid line: 10 step ahead predicted pitch rate, based on the fourth order model from canard angle only.

Using all inputs

u_1 canard angle; u_2 Elevator angle; u_3 Leading edge flap;

$$\begin{aligned} y(t) = & -a_1 y(t-T) - a_2 y(t-2T) - a_3 y(t-3T) - a_4 y(t-4T) \\ & +b_1^1 u_1(t-T) + \ldots + b_1^4 u_1(t-4T) \\ & +b_2^1 u_2(t-T) + \ldots + b_1^3 u_3(t-T) + \ldots + b_4^3 u_3(t-4T) \end{aligned}$$



System Identification: Issues

→ 14

- Select a class of candidate models
- Select a member in this class using the observed data
- Evaluate the quality of the obtained model
- Design the experiment so that the model will be “good”.

System Identification: State-of-the-Art Setup

A Typical Problem

Given Observed Input-Output Data: Find a Description of the System that Generated the Data [Simulator or Predictor. Linear System: Impulse response or Bode plot].

Basic Approach

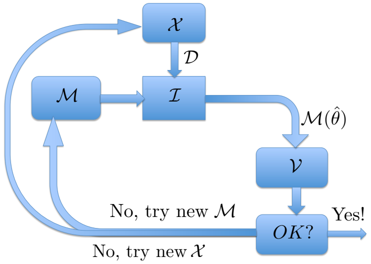
Find a suitable Model Structure, Estimate its parameters, and compute the response of the resulting model

Techniques

Estimate the parameters by ML techniques/PEM (prediction error methods). Find the model structure by AIC, BIC or Cross Validation

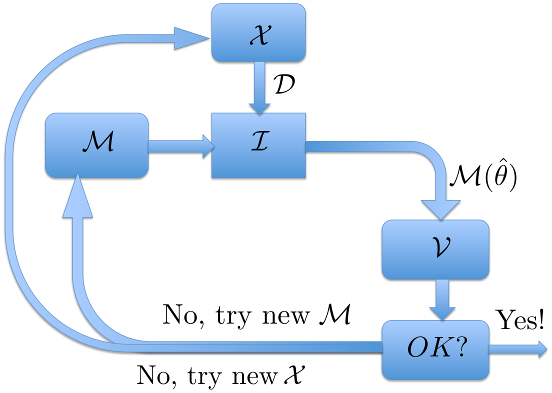
The SI Flow

$\rightarrow$ 15


$\mathcal{X}$
$\mathcal{D}$
$\mathcal{M}$
$\mathcal{I}$
$\mathcal{M}(\hat{\theta})$
$\mathcal{V}$
$OK?$
Yes!
No, try new $\mathcal{M}$
No, try new $\mathcal{X}$

$\mathcal{X}$: The Experiment
$\mathcal{D}$: The Measured Data
$\mathcal{M}$: The Model Set
$\mathcal{I}$: The Identification Method
$\mathcal{V}$: The Validation Procedure

The SI Flow; Model Structures $\mathcal{M}$


$\mathcal{X}$
$\mathcal{D}$
$\mathcal{M}$
$\mathcal{I}$
$\mathcal{M}(\hat{\theta})$
$\mathcal{V}$
$OK?$
Yes!
No, try new $\mathcal{M}$
No, try new $\mathcal{X}$

Models: General Aspects

$\rightarrow$ 79,161

- A model is a mathematical expression that describes the connections between measured inputs and outputs, and possibly related noise sequences.
- They can come in many different forms
- The models are labeled with a parameter vector θ
- A common framework is to describe the model as a predictor of the next output, based on observations of past input-output data.
  Observed input–output (u, y) data up to time t: Z^t
  Model described by predictor: $\mathcal{M}(\theta) : \hat{y}(t|\theta) = g(t, \theta, Z^{t-1})$.

Linear Models

$\rightarrow$ 79

General Description

$y(t) = G(q, \theta)u(t) + H(q, \theta)e(t)$, q : shift op. e : white noise

$$G(q, \theta)u(t) = \sum_{k=1}^{\infty} g_k u(t-k), \quad H(q, \theta)e(t) = 1 + \sum_{k=1}^{\infty} h_k e(t-k)$$

$$y(t) = G(q, \theta)u(t) + v(t); \quad \text{Spectrum } \Phi_v(\omega) = \lambda |H(e^{i\omega})|^2$$

Predictor

$$\hat{y}(t|\theta) = G(q, \theta)u(t) + [I - H^{-1}(q, \theta)][y(t) - G(q, \theta)u(t)]$$

Common Black-Box Parameterizations

→ 81

BJ (Box-Jenkins)

$$G(q,\theta) = \frac{B(q)}{F(q)}; \quad H(q,\theta) = \frac{C(q)}{D(q)}$$

$$B(q) = b_1 q^{-1} + b_2 q^{-2} + \ldots b_{nb} q^{-nb}$$

$$F(q) = 1 + f_1 q^{-1} + \ldots + f_{nf} q^{-nf}$$

$$\theta = [b_1, b_2, \ldots, f_{nf}]$$

ARX:

$$y(t) = \frac{B(q)}{A(q)} u(t) + \frac{1}{A(q)} e(t) \text{ or}$$

$$A(q)y(t) = B(q)u(t) + e(t) \text{ or}$$

$$\begin{aligned} y(t) + a_1 y(t-1) + \ldots + a_{na} y(t-na) \\ = b_1 u(t-1) + \ldots + b_{nb} u(t-nb) \end{aligned}$$

Special Feature of ARX-models

The ARX Model is a Linear Regression

The predictor for ARX can be written

$$\hat{y}(t|\theta) = \varphi^T(t)\theta$$

$$\varphi(t) = \begin{bmatrix} -y(t-1) & \ldots & -y(t-na) & u(t-1) & \ldots & u(t-nb) \end{bmatrix}^T$$

$$\theta = \begin{bmatrix} a_1 & \ldots & a_{na} & b_1 & \ldots & b_{nb} \end{bmatrix}^T$$

Common Black and Grey Parameterizations

→ 97

State-Space with Possibly Physically Parameterized Matrices

$$\begin{aligned} x(t+1) &= A(\theta)x(t) + B(\theta)u(t) + K(\theta)e(t) \\ y(t) &= C(\theta)x(t) + e(t) \end{aligned}$$

Corresponds to

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

$$H(q,\theta) = C(\theta)(qI - A(\theta))^{-1}K(\theta) + I$$

Continuous Time (CT) Models

→ 93

Physical Model with unknown parameters

$$\begin{aligned} \dot{x}(t) &= \mathcal{F}(\theta)x(t) + \mathcal{G}(\theta)u(t) + w(t) \\ y(t) &= C(\theta)x(t) + D(\theta)u(t) + v(t) \end{aligned}$$

Sample it (with correct Input Intersample Behaviour):

$$\begin{aligned} x(t+1) &= A(\theta)x(t) + B(\theta)u(t) + K(\theta)e(t) \\ y(t) &= C(\theta)x(t) + e(t) \end{aligned}$$

Now apply the discrete time formalism to this model, which is parameterized in terms of the CT parameters θ

Non-Linear Models

$\boxed{\rightarrow 146}$

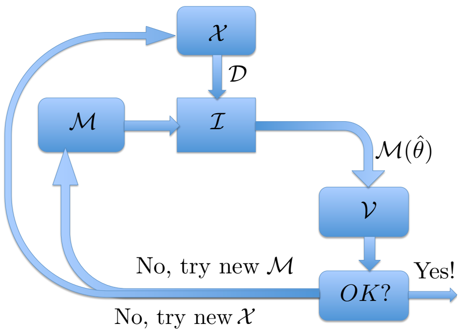
$\mathcal{M}(\theta) : \hat{y}(t|\theta) = g(t, \theta, Z^{t-1})$. g non-linear in Z.

- More in the NL lecture notes
- Physical Grey-Box Models
  Perform physical modeling (e.g. in MODELICA) and denote unknown physical parameters by θ. Collect the model equations as

$$\begin{aligned} \dot{x}(t) &= f(x(t), u(t), \theta) \\ y(t) &= h(x(t), u(t), \theta) \end{aligned}$$

- NLARX models
- Block Oriented Models, Wiener-Hammerstein and similar
- Local Linear Models, Linear Parameter Varying Models.
  Linear model $G(z, p)$ parameterized by a measured regime variable p

The SI Flow: Estimation: $\mathcal{I}$


$\mathcal{X}$
$\mathcal{D}$
$\mathcal{M}$
$\mathcal{I}$
$\mathcal{M}(\hat{\theta})$
$\mathcal{V}$
No, try new $\mathcal{M}$
Yes!
$OK?$
No, try new $\mathcal{X}$

Criteria: 1. Pragmatically

$\boxed{\rightarrow 199}$

If a model, $\hat{y}(t|\theta)$, essentially is a predictor of the next output, is is natural to evaluate its quality by assessing how well it predicts: Form the *Prediction error* and measure its size:

$$\varepsilon(t, \theta) = y(t) - \hat{y}(t|\theta), \quad \ell(\varepsilon(t, \theta))$$

(Note: Linear system $\varepsilon = \frac{1}{H}(y - Gu)$) Typically $\ell(x) = x^2$. How has it performed historically?

$$V_N(\theta) = \sum_{t=1}^{N} \ell(\varepsilon(t, \theta))$$

Which model in the structure performed best?

$$\hat{\theta}_N = \arg \min_{\theta \in D_{\mathcal{M}}} V_N(\theta)$$

Criteria: 2. Statistically

$\boxed{\rightarrow 212}$

Maximum Likelihood (ML): Which model makes the recorded observations most likely? (pdf: probability density function)

$\max p(Y^N|\theta)$ $\quad p$: the pdf of the outputs for a given parameter value

If the innovations e of the measured data have the pdf $f(x)$, then

$$-\log p(Y^N|\theta) = V_N(\theta) = \sum_{t=1}^{N} \ell(\varepsilon(t, \theta)) \qquad \ell(x) = -\log f(x)$$

So, what minimizes the ”pragmatic fit” is also the ML estimate!
Gaussian pdf ⇒ $\ell(x) \sim x^2$

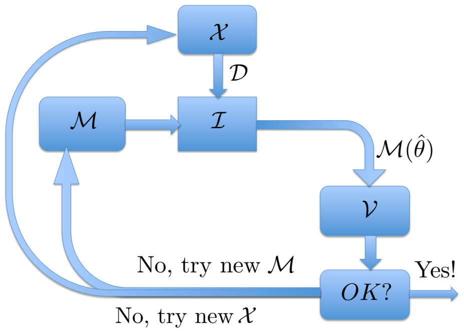
Estimation: Variants

- More in the Special Issues Lecture notes
- **Regularization:** To curb the flexibility of the model structure and to provide better numerics in the minimization we can postulate a regularized criterion: → 221

$$W_N(\theta) = V_N(\theta) + \lambda(\theta - \theta^\dagger)^T R((\theta - \theta\dagger)$$

- **Bayesian View:** The parameter vector θ is a random vector with a certain prior pdf, and we seek the posterior pdf, given the observations. Suppose the prior is $\theta \in N(\theta^\dagger, R^{-1}/\lambda)$ → 213
- Sub-space methods → 208
- IV-techniques → 224

Model Estimate Properties $\mathcal{M}(\hat{\theta}_N)$


$\mathcal{X}$
$\mathcal{D}$
$\mathcal{M}$
$\mathcal{I}$
$\mathcal{M}(\hat{\theta})$
$\mathcal{V}$
No, try new $\mathcal{M}$
$OK?$
Yes!
No, try new $\mathcal{X}$

Model Estimate Properties, → 223, 287

As the number of data, N, tends to infinity

- $\hat{\theta}_N \to \theta^* \sim \arg\min_\theta E\ell(\varepsilon(t,\theta))$ the best possible predictor in $\mathcal{M}$
- If $\mathcal{M}$ contains a true description of the system
  - Cov $\hat{\theta}_N = \frac{\lambda}{N}[E\psi(t)\psi^T(t)]^{-1}$ [$\psi(t) = \frac{d}{d\theta}\hat{y}(t|\theta), \lambda$: noise level]...
  - ... is the Cramér-Rao lower bound for any (unbiased) estimator.

E: Expectation. These are very nice optimal properties:

- The model structure is large enough: The ML/PEM estimated model is (asymptotically) the best possible one. Has smallest possible variance (Cramér- Rao)
- The model structure is not large enough: The ML/PEM estimate converges to the best possible approximation of the system. ”The estimate has smallest possible asymptotic bias.”

Proof: Formal Calculations

We illustrate the main ideas for a simple curve-fitting problem:
System: $y(t) = g_0(x_t) + e(t)$, e white noise. Model: $g(x_t, \theta)$
Convergence: (LLN= Law of Large Numbers)

$$\begin{aligned} V_N(\theta) &= \frac{1}{N}\sum(g_0(x_t) + e(t) - g(x_t,\theta))^2 \\ &= \frac{1}{N}\sum(g_0(x_t) - g(x_t,\theta))^2 + \frac{1}{N}\sum e^2(t) \\ &+ \frac{2}{N}\sum(g_0(x_t) - g(x_t,\theta))e(t) \end{aligned}$$

LLN: $\frac{1}{N}\sum(g_0(x_t) - g(x_t,\theta))e(t) \to 0$ (uniformly in θ!)

so $V_N(\theta) \to \bar{E}(g_0(x_t) - g(x_t,\theta))^2$

$= \lim \frac{1}{N}\sum(g_0(x_t) - g(x_t,\theta))^2$ as $N \to \infty$

Proof: Formal Calculations, ctn'd

Asymptotic Distribution (CLT= Central Limit Theorem)

$$0 = V_N'(\hat{\theta}_N) = V_N'(\theta^*) + V_N''(\theta^*)(\hat{\theta}_N - \theta^*)$$

$$(\hat{\theta}_N - \theta^*) = -[V_N''(\theta^*)]^{-1} V_N'(\theta^*)$$

$$V_N'(\theta) = \frac{2}{N}\sum (y(t) - g(x_t,\theta))g'(x_t,\theta)$$

$$V_N'(\theta^*) = \frac{2}{N}\sum e(t)\psi(t); \quad \psi(t) = g'(x_t;\theta^*)$$

$$\text{LLN: } V_N''(\theta^*) = \frac{2}{N}\sum \psi(t)\psi^T(t) + \frac{2}{N}\sum e(t)g''(x_t,\theta^*) \to 2\bar{E}\psi\psi^T$$

$$\text{CLT: } \frac{1}{\sqrt{N}}\sum e(t)\psi(t) \to N(0,\lambda\bar{E}\psi\psi^T)$$

$$\sqrt{N}(\hat{\theta}_N - \theta^*) \to N(0,\lambda[\bar{E}\psi\psi^T]^{-1})$$

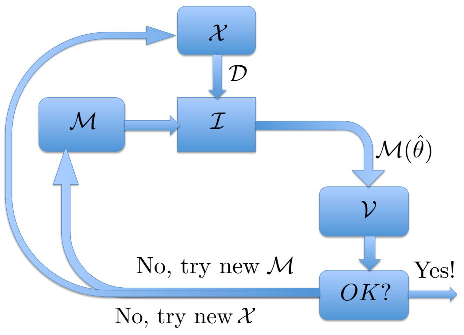
Asymptotics for linear models $\boxed{\to 263, 295}$

[Φ_u, Φ_v: Spectra of input and additive noise $v = He$. G_0 true transfer function.]

$$\hat{\theta}_N \to \theta^* = \arg\min_\theta \int_{-\pi}^{\pi} |G(e^{i\omega},\theta) - G_0(e^{i\omega})|^2 \frac{\Phi_u(\omega)}{|H(e^{i\omega},\theta)|^2} d\omega$$

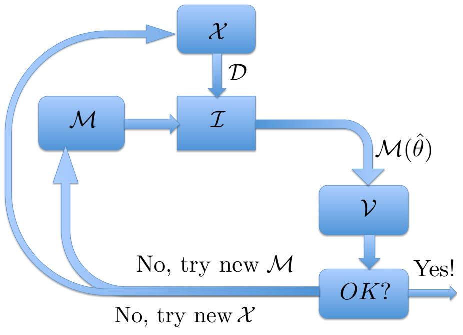
$$\text{Cov}G(e^{i\omega},\hat{\theta}_N) \sim \frac{n}{N}\frac{\Phi_v(\omega)}{\Phi_u(\omega)} \text{ as } n, N \to \infty \quad n: \text{model order}$$

Experiment Design: $\mathcal{X}$


$\mathcal{X}$
$\mathcal{D}$
$\mathcal{M}$
$\mathcal{I}$
$\mathcal{M}(\hat{\theta})$
$\mathcal{V}$
No, try new $\mathcal{M}$
Yes!
$OK?$
No, try new $\mathcal{X}$

See Lecture Notes "Experiment Design"

Model Validation:$\mathcal{V}$


$\mathcal{X}$
$\mathcal{D}$
$\mathcal{M}$
$\mathcal{I}$
$\mathcal{M}(\hat{\theta})$
$\mathcal{V}$
No, try new $\mathcal{M}$
Yes!
$OK?$
No, try new $\mathcal{X}$

Model Validation Techniques

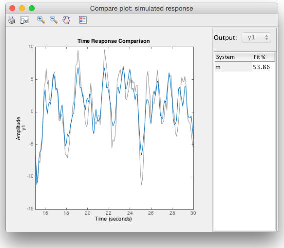
“Twist and turn” the model(s) to check if they are good enough for the intended application.
Essentially a subjective decision. Several basic techniques are available:

1. Simulation (prediction) cross validation
   - Check how well the estimated model can reproduce new, validation data
2. Residual Analysis
   - Are the "leftovers" unpredictable?
3. Comparing Different Models
   - Which one performs best?
4. Bias-Variance Trade-off
   - What is the suitable complexity (flexibility) of a model?

Simulation Cross Validation

Collect a *estimation data set* $\mathcal{Z}_e$ and a *validation data set* $\mathcal{Z}_v$.
Estimate the model(s) using the estimation data and simulate it using the input signal in the validation data. Plot that simultated model output together with the measured validation output.

```
m = arx(z1(1:150),[2 2 1]);
compare(z1(151:end),m)
```



Output: y1

System	Fit %
m	53.86

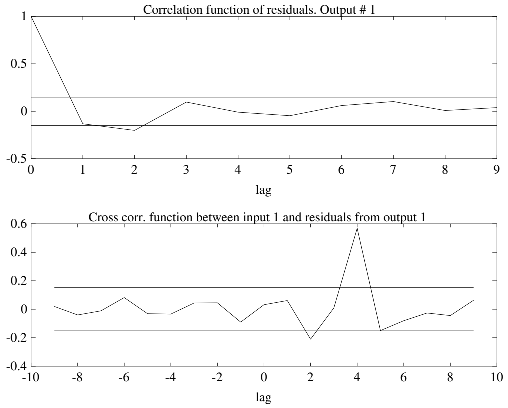
“The ultimate test”. Needs no probabilistic justification.
Plot simulated outputs and measured outputs for the fresh data set. Alternatively one may use predictions over longer periods.
Typically the performance can be evaluated as sum of squared mismatches. (“FIT”)

Model Validation: Residual Analysis

$$\varepsilon(t) = \varepsilon(t, \hat{\theta}_N) = y(t) - \hat{y}(t|\hat{\theta}_N)$$

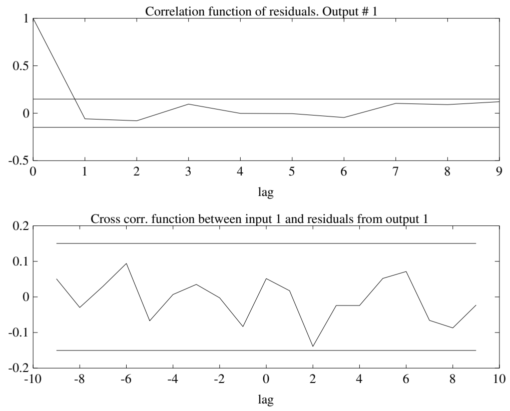
- $\varepsilon(t)$ should be independent of $u(t-\tau), \tau > 0$
- $\varepsilon(t)$ should ideally be white noise.
- Compute the corresponding correlation functions
- Confidence intervals for the estimates can also be computed. This requires some care!

Typical Plots



True delay: 4 guessed delay: 5

Typical Plots, ct'd



Guessed delay = true delay.

Model Validation Techniques

→ 493

"Twist and turn" the model(s) to check if they are good enough for the intended application.
Essentially a subjective decision. Several basic techniques are available:

1. Simulation (prediction) cross validation
   - Check how well the estimated model can reproduce new, validation data
2. Residual Analysis
   - Are the "leftovers" unpredictable?
3. Comparing Different Models
   - Which one performs best?
4. Bias-Variance Trade-off
   - What is the suitable complexity (flexibility) of a model?

Comparing Different Models

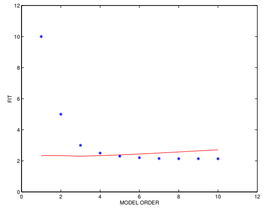
→ 498

- What to compare?
  - simulation performance
  - prediction performance
- `compare(zv,th,k)`
- Comparing models on fresh data sets (Simulation cross validation as before)
- Comparing models on second-hand data sets

Comparing Models on Second-Hand Data

→ 501

Problem: Larger models will always perform better on estimation data.
BASIC IDEA: Compensate for over-fit



- Add Model Complexity Penalty
  - Akaike: AIC, FPE
  - Rissanen MDL
- Hypothesis test
  - Check if decrease in fit is larger than "expected"

AIC, FPE, MDL

→ 503

AIC for Gaussian case:

$$\min_{\mathcal{M}} \min_{\theta_{\mathcal{M}}} [(1+\frac{2d_{\mathcal{M}}}{N}) \cdot \frac{1}{N}\sum_{t=1}^{N} \varepsilon^2(t,\theta_{\mathcal{M}})], \quad d_{\mathcal{M}} = \dim \theta_{\mathcal{M}}$$

[-log likelihood + $d_{\mathcal{M}}$]

FPE similar, with $(1+\frac{2d_{\mathcal{M}}}{N})$ replaced by $\frac{N+d_{\mathcal{M}}}{N-d_{\mathcal{M}}}$. Aims at estimating the fit that would be obtained for a fresh data set.

MDL (or BIC) is like AIC but with a $\frac{2d_{\mathcal{M}}\log N}{N}$ penalty.

Comparing Different Models

→ 498

- What to compare?
  - simulation performance
  - prediction performance
- `compare(zv,th,k)`
- Comparing models on fresh data sets (Simulation cross validation as before)
- Comparing models on second-hand data sets
- Comparing many models simulaneously

Bias – Variance Trade Off

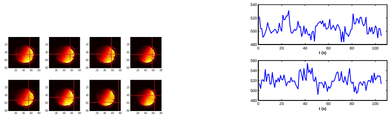
→ 492

Any estimated model is incorrect. The errors have two sources:

- Bias: The model structure is not flexible enough to contain a correct description of the system.
- Variance: The disturbances on the measurements affect the model estimate, and cause variations when the experiment is repeated, even with the same input.

Mean Square Error (MSE) = $|\text{Bias}|^2$ + Variance.
When model flexibility ↑,Bias ↓ and Variance ↑.
To minimize MSE is a good trade-off in flexibility.
In state-of-the-art Identification, this flexibility trade-off is governed primarily by model order.

Status of the State-of-the-Art Framework

- Well established statistical theory
- Optimal asymptotic properties
- Efficient software
- Many applications in very diverse areas. Some examples:
  - Aircraft Dynamics:
  - Brain Activity (fMRI): 
  - Pulp Buffer Vessel: ;