System Identification, Lecture 6

Kristiaan Pelckmans (IT/UU, 2338)

Course code: 1RT875, Report code: 61806,
F, FRI Uppsala University, Information Technology

03 Februari 2010

Lecture 6

- Model Structure Determination and Model Validation (Ch. 11)

"A model is no use if its validity is not verified."

Choice of Model Structure

- Type of model set.
  - Ex.: Linear, Nonlinear, black- or white box models
  - Here: ARX, OE, ARMAX, ...
- Size of the model set. Orders of the polynomials ($A(q^{-1})$, $B(q^{-1})$, $C(q^{-1})$, $D(q^{-1})$, ...). Not the 'true' orders in reality..
- Model Parametrization:
  - Transformation of data.
  - Choice of operators: e.g. $q \leftrightarrow \frac{q-1}{h}$.

Objective: Obtain a good model at a low cost!

- Quality of model: A scalar measure of goodness, e.g. the mean square error (MSE)
  - MSE consists of bias+variance:

$$\mathrm{MSE}(\hat{\theta}) = E\|\hat{\theta} - \theta_0\|^2 = E\|E[\hat{\theta}] - \hat{\theta}\|^2 + E\|E[\hat{\theta}] - \theta_0\|^2$$

  - Reduce bias $\rightarrow$ more flexible model structures.
  - Decrease variance $\rightarrow$ decrease the number of estimated parameters.
  - Trade-off: flexibility versus parsimony.
- Price of modeling:
  - Algorithm complexity.
  - Computation time.
- Intended use of the model!

Model Validation

Reasons:

- Underparametrized.
- Overparametrized.

Basic Approaches:

- Plots of signals.
- Common sense (will the model serve its purpose?)
- Statistical 'goodness of fit' tests.

Basic Plots and Common Sense

- Compare the measured output with the simulated output

$$\hat{y} = G(q^{-1}, \hat{\theta}_N)u(t).$$

  The differences $y(t) - \hat{y}$ (not prediction errors!) are due to disturbances and modeling errors.

- Plot the differences $y(t) - \hat{y}$.

- Compare a step response of the system to the step response of the model.

- Compare the nonparametric estimate of the transfer function to the transfer function of the model (frequency model).

  Figure ... .

What to compare?

Def. The k-step ahead model predictors $\hat{y}(t, \hat{\theta}|t-k)$ are based on the data

$$u(1), \ldots, u(t-k), y(1), \ldots, y(t-k),$$

using the estimate $\hat{\theta}$

Common choice are

- $\hat{y}(t, \hat{\theta}|t-1)$ is the mean square optimal predictor

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

- $\hat{y}_\infty(t, \hat{\theta})$, only based on past inputs (referred to as simulation)

$$\hat{y}(t, \hat{\theta}) = G(q^{-1}, \hat{\theta})u(t)$$

To compare different models, we use a scalar measure, e.g.

$$V^k(\hat{\theta}) = \frac{1}{N}\sum_{t=1}^{N}\Big(y(t) - \hat{y}(t, \hat{\theta}|t-k)\Big)^2$$

Example. Use $\hat{y}(t, \hat{\theta}|t-k) = y(t-k)$.

Question

In the following we will concern ourselves with the following questions:

- Is the model structure flexible enough to cover the observed behavior?
- Is a given model too complex?
- Given two different models, which one should be chosen?

Is a model Flexible enough?

The 'leftovers' from the modeling process - the part of the data that the model could not reproduce - are 'the residuals'

$$\epsilon(t) = \epsilon(t, \hat{\theta}) = (y(t) - \hat{y}(t, \hat{\theta}|t-1))^2.$$

Rem. the residuals are the prediction errors evaluated at $\hat{\theta}$. If $\hat{\theta} = \theta_0$, the residuals are white.

- If

$$\hat{\mathbf{R}}_\epsilon(\tau) = \frac{1}{N} \sum_{t=1}^{N-\tau} \epsilon(t)\epsilon(t+\tau)^T$$

  is not small for $\tau \neq 0$, then part of $\epsilon(t)$ could be 'explained' from past data. This means that $y(t)$ could have been predicted better.

- The cross-covariance between residuals and inputs

$$\hat{\mathbf{R}}_{\epsilon,u}(\tau) = \frac{1}{N}\sum_{t=1}^{N-\tau} \epsilon(t)u(t+\tau)^T$$

should be small if the model has picked up the essential part of the dynamics from u to y (assuming open loop identification). This also indicates that the residual test should be invariant to inputs.

Testing Whiteness

If the model is accurately describing the observed data, then the residuals $\epsilon(t)$ should be white. A way to verify this is to test for the hypothesis.

$$\begin{cases} H_0: & \epsilon(t) \text{ is white} \\ H_1: & \text{otherwise} \end{cases}$$

this can be done in several ways, for example:

Autocorrelation Test

The autocorrelation of the residuals (single output) are estimated as

$$\hat{r}_\epsilon(\tau) = \frac{1}{N} \sum_{t=1}^{N-\tau} \epsilon(t)\epsilon(t+\tau)$$

If H_0 holds, then the squared covariance should be χ^2-distributed, or

$$\frac{N}{\hat{r}^2(0)} \sum_{i=1}^{m} \hat{r}^2(i) \to \chi^2(m).$$

Furthermore, the normalized auto-covariance estimates are asymptotically Gaussian distributed, or for all $\tau = 1, \ldots, m$

$$\sqrt{N} \frac{\hat{r}^2(\tau)}{\hat{r}^2(0)} \to \mathcal{N}(0, 1).$$

A typical way to use the first test statistic is as follows (the second can be used similarly). Let x be a random variable which is distributed as $\chi^2(m)$, define χ^2_α for given α as follows

$$\alpha = P\left(x > \chi^2_\alpha\right).$$

Then for some $\alpha = 0.1, 0.01$ one has

$$\begin{cases} \frac{N}{\hat{r}^2(0)} \sum_{i=1}^m \hat{r}^2(i) > \chi^2_\alpha & \text{reject} \quad H_0 \\ \frac{N}{\hat{r}^2(0)} \sum_{i=1}^m \hat{r}^2(i) \leq \chi^2_\alpha & \text{accept} \quad H_0 \end{cases}$$

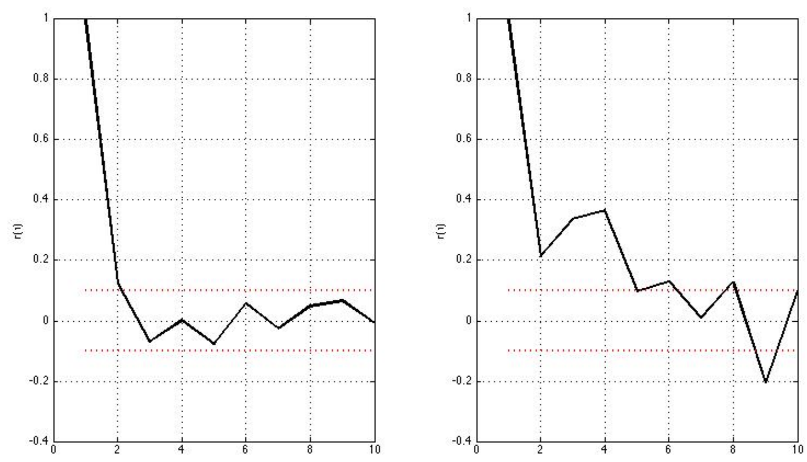

Figure 1: test of the autocorrelation sequence (a) accept H_0, (b) reject H_0.

Zero Crossing Test

Given a white noise sequence, one can expect that the residuals change sign on the average every second time step. Introduce x_N as the number of times the residual changes sign up to moment N, or

$$x_N = \sum_{t=1}^{N-1} I(\epsilon(t)\epsilon(t+1) < 0)$$

then it can be shown that

$$x_N \rightarrow \mathcal{N}(m, p),$$

where $m \approx N/2$ and $p \approx N/4$, or

$$\frac{2x_N - N}{\sqrt{N}} \rightarrow \mathcal{N}(0, 1).$$

Cross-Correlation Test

If the model as an accurate description of the system, then the input and residuals should be uncorrelated (no unmodeled dynamics), or

$$\hat{r}_{\epsilon,u}(\tau) = \frac{1}{N} \sum_{t=1}^{N-\tau} \epsilon(t) u(t+\tau) \to 0$$

- If $\lim_{N \to \infty} \hat{r}_{\epsilon,u}(\tau) \neq 0$, then there is output feedback in the input
- ... or indicating wrong time-delay in the model. If a time delay of two samples has been assumed in the model, but the 'true' time delay is 1, clearly $E[u(t-1)\epsilon(t)] \neq 0$
- This can be seen by rewriting the model as

$$\epsilon(t) = G'(q^{-1}) u(t)$$

The following result can be used to design a hypothesis test if inputs/residuals are uncorrelated. Let

$$\hat{\mathbf{R}}_u = \frac{1}{N}\sum_{t=m+1}^{N} \begin{bmatrix} u(t-1) \\ \vdots \\ u(t-m) \end{bmatrix} \begin{bmatrix} u(t-1) & \dots & u(t-m) \end{bmatrix}$$

and

$$\hat{r}_m = \begin{bmatrix} \hat{r}_{u\epsilon}(\tau+1) & \dots & \hat{r}_{u\epsilon}(\tau+1) \end{bmatrix}^T$$

then

$$N\hat{r}_m^T \left(\hat{r}_\epsilon(0)\hat{\mathbf{R}}_u\right)^{-1} \hat{r}_m \to \chi^2(m)$$

can be used to design a hypothesis test.

Is a model too complex?

If a model is *overparameterized* , it is unneccesarily complicated and can be sensitive to parameter variations. For example, look at the pole-zero behavior of the estimated transfer function in such a case.

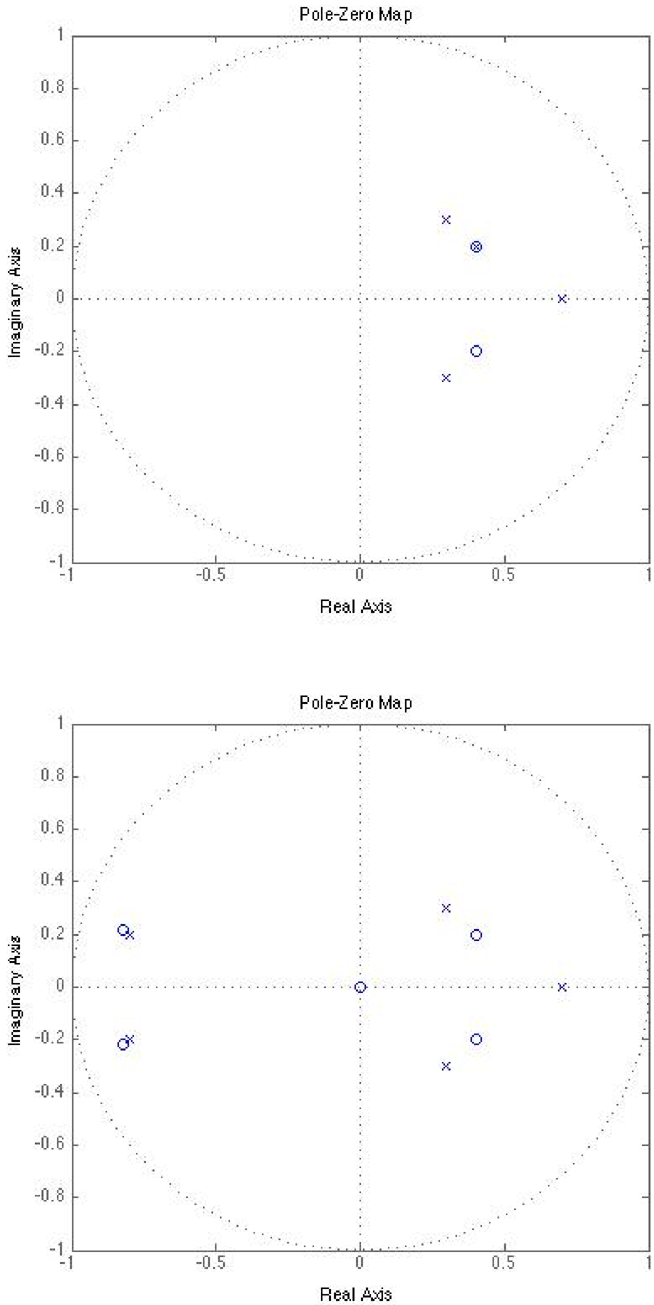

Figure 2: Pole-Zero Cancelation: (i) ARX (3,2), (ii) ARX(5,4)

Cross-validation

- Check the criterion $V^1(\hat{\theta})$. A model structure that is too large will also model the disturbances in the given data. This is called *overfitting* the data.

- Using a 'fresh' dataset that was not used during identification, is called 'validation'.

- (Cross-) validation is a nice and simple way to compare models to detect 'overfitted' models.

- (Cross-) validation requires a large amount of data, as the 'validation' data cannot be used during identification.

The Parsimony Principle

'Simple models are to be preferred (Occam's razor)' Assume a model quality is measures by $E[V^1(\hat{\theta})]$

- If $\hat{\theta} = \theta_0$, the residuals would equal the noise and $E[V^1(\theta_0)] = \lambda^2$.
- In the Least Squares case, we had tat $E[V^1(\hat{\theta})] = \frac{N-n}{N}\lambda^2$

Hence, each extra parameter will 'bias' the criterion a factor $\frac{\lambda^2}{N}$...

Comparisons of Model Structures

Use the PEM loss function $V^1(\hat{\theta})$ as a measure of the model quality. For models of increasnig order, the value of this loss will decrease monotonically, and the problem is to find the lowest model order that gives acceptable loss.

Let V and V' be the loss of two models for two different model orders n and n'. Then

$$x_N \triangleq N \frac{V^2 - {V'}^2}{V^2} \to \chi^2(n' - n).$$

Hence we choose model order n at significance level α if

$$x_N \leq \chi^2_\alpha(n' - n),$$

otherwise select n'.

Another approach is to formulate a criterion that is a function of the loss $V^1(\hat{\theta})$, and which also penalizes the model order:

$$\mathrm{IC}(\mathcal{M}) = V^1(\hat{\theta})(1 + \beta(N, n))$$

where $\beta(N, n)$ is a function increasing proportional to n, but decreasing (to zero when $N \to 0$) proportional to N.

Important examples of penalization functions are

1. (Akaike AIC)

$$\mathrm{AIC}(\mathcal{M}) \propto \log V^1(\hat{\theta}) + \frac{2n}{N}$$

2. (Final Prediction Error FPE)

$$\mathrm{FPE}(\mathcal{M}) = V^1(\hat{\theta}) \left(\frac{1 + n/N}{1 - n/N} \right)$$

3. (Minimum Description Length, MDL)

$$\mathrm{MDL}(\mathcal{M}) = V^1(\hat{\theta}) \left(1 + \frac{n \log(N)}{N} \right)$$

The AIC and the FPE are asymptotically equivalent, but it can be shown that both will tend to select to high model orders. The

MDL yields consistent estimates. Again, physical insight might significantly help the analysis.

Summary - Model Validation

- Many different tests can be used to verify the validity of a model (try simple things first).
- Th choice of an appropriate model structure (model order) can be based on a statistical test based on the residuals (auto- and cross-correlation tests).
- To decide on the appropriate model order, AIC, FPE or MDL can be used.
- (Cross-) validation is best if lots of data is available.
- Implementations available in the MATLAB SI toolbox.