System Identification, Lecture 6

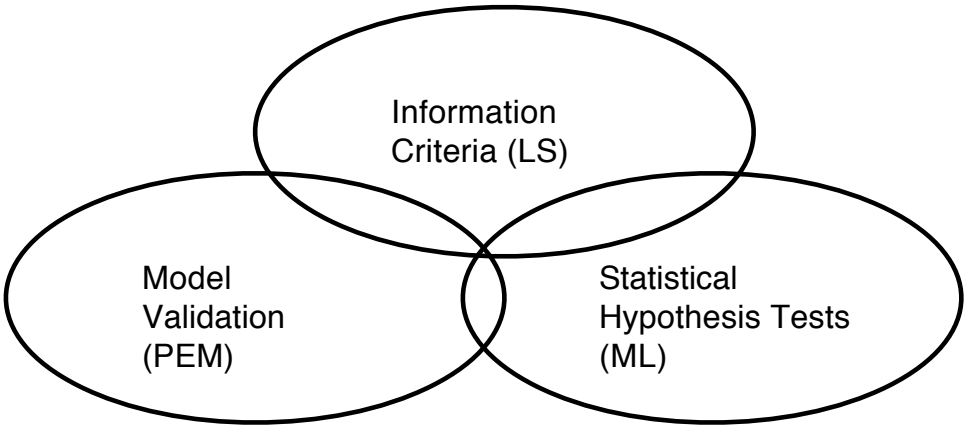
Kristiaan Pelckmans (IT/UU, 2338)

Course code: 1RT885, Report code: 61800 - Spring 2017
F, FRI Uppsala University, Information Technology

26 April 2017

Model Validation and Model Structure Determination

- What?
- Common Sense.
- Model Validation Procedures.
- Statistical Hypothesis Testing.
- Information Criteria.



Q: If we have an estimate using PEM, why is a whiteness test of the residuals a good/bad idea?

What?

Many choices to verify:

- Before estimation: model structure (ARX, OE, ...)?
- Orders of polynomials $A(q^{-1}), B(q^{-1}), \ldots$.
- Transformation of data?
- Choice of preprocessing (sampling, delay,...)?
- After modeling: sufficient for our need?
- Inherent difficulties to the estimate?
- Identification vs. approximation.

Objective:

- Quality of the model for intended use.
- Computational Cost.
- Numerical Cost.
- Algorithmic Cost.
- Trade-offs: bias vs. variance.

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

What goes wrong?

- Under-parametrized.
- Over-parametrized.
- High Variance:

$$\mathbb{E}[V_n(\theta_n)] = \left(\frac{n-d}{n}\right)\lambda^2$$

- Data not sufficiently rich (PE).
- Unidentifiable components.

$$y_t = \frac{(1-aq^{-1})(1-cq^{-1})}{(1-bq^{-1})(1-cq^{-1})}u_t$$

- Assumptions for estimation not valid.
- Not LTI.

Common Sense

- Compare the measured output with the simulated output

$$\hat{y}_t = H(q^{-1}, \theta_n) u_t.$$

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

- Plot the differences $y_t - \hat{y}_t$.

- 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).

- Implement candidate model.

Pole-zero Cancellation:

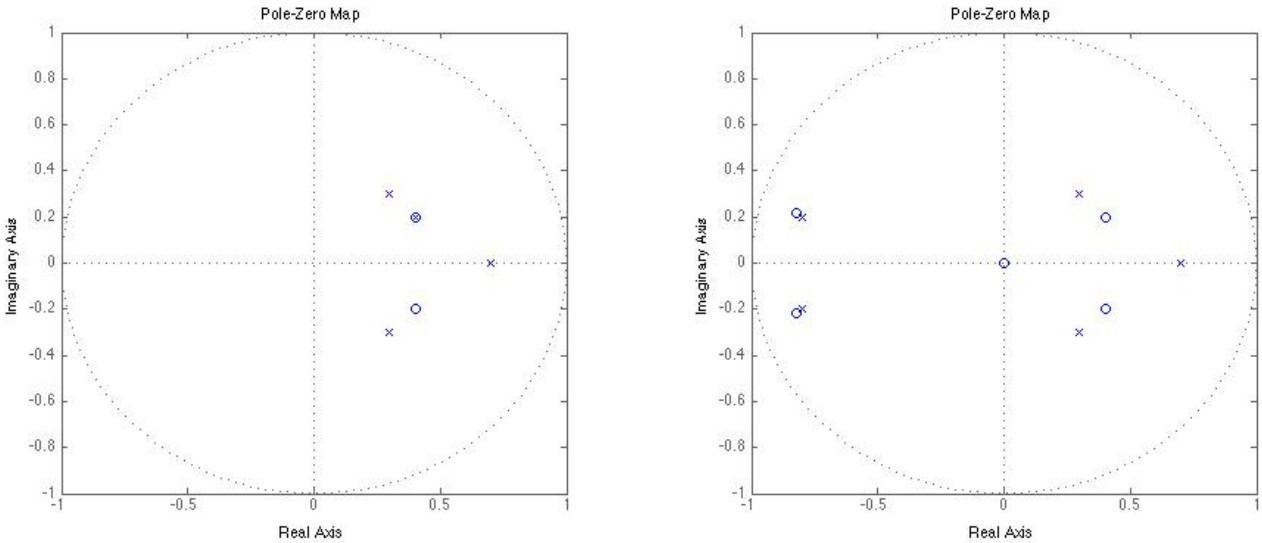


Figure 1: Pole-Zero Cancelation: (i) ARX (3,2), (ii) ARX(5,4)

Model validation Procedures

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

$$u_1, \ldots, u_{t-k}, y_1, \ldots, y_{t-k},$$

using the estimate θ_n

Common choice are

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

$$\hat{y}_t(\theta_n | t-1) = G^{-1}(q^{-1}, \theta_n) H(q^{-1}, \theta_n) u_t + (1 - G^{-1}(q^{-1}, \theta_n)) y_t$$

- $\hat{y}_t^{\infty}(\theta_n)$, only based on past inputs (referred to as simulation)

$$\hat{y}_t(\theta_n) = H(q^{-1}, \theta_n) u_t$$

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

$$V_n^k(\theta_n) = \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \hat{y}_t(\theta_n|t-k)\right)^2$$

Example. Use $\hat{y}_t(\theta_n|t-k) = y_{t-k}$.

- Validation.
- Compare Prediction performance on train-validation set ('overfitting').
- Cross-validation.
- Leave-one-out Cross-validation.
- Large timeseries.
- Jack-knife and Bootstrap.
- Monte-Carlo sampling.

Statistical Hypothesis Testing

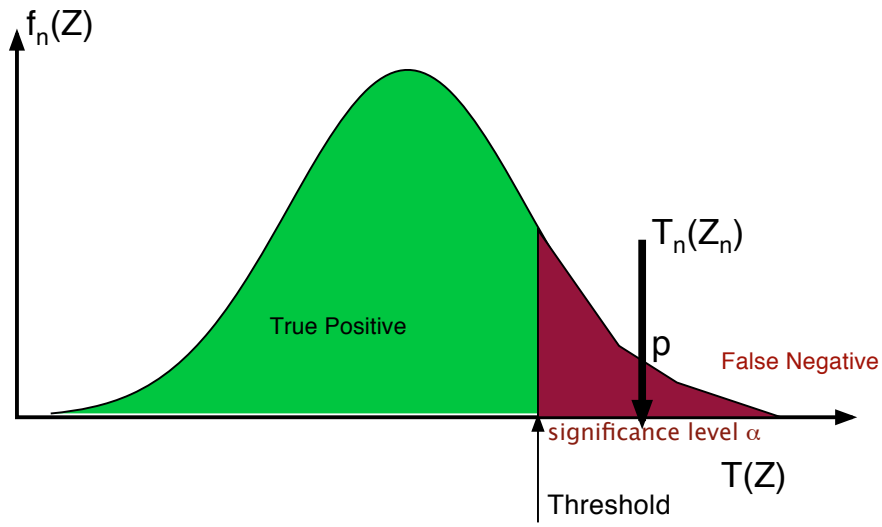
Assuming stochastic setup with r.v. $\{D_t\}_t$ and/or $\{U_t\}_t$:

- Test if assumptions made during estimation valid.
- Test if prior knowledge valid.
- Test if data 'fits in' assumed model structure.
- Test if the model is too complex.
- Test which of two models is more performant.

How does it work?

- First define $\mathcal{H}_0$ and $\mathcal{H}_1$.
- Then test under which hypothesis the data is more likely.
- Or accept/reject $\mathcal{H}_0$.
- Devise a 'statistic' $T(Z)$.
- Derive the 'theoretical' sample distribution f_T of $T(Z)$ under $Z \in \mathcal{H}_0$.
- If $T(Z_n)$ is not likely to be sampled from this sample distribution, then reject with certain power $0 < \alpha$.

$$\mathcal{H}_0 : \; f_T\left(T(Z_n)\right) = \boxed{p} \leq t_\alpha.$$



WIKI - ... in a proposed experiment to test a Lady's claimed ability to determine the means of tea preparation by taste.

1. The null hypothesis was that the Lady had no such ability.
2. The test statistic was a simple count of the number of successes in 8 trials.
3. The distribution associated with the null hypothesis was the binomial distribution familiar from coin flipping experiments.
4. The critical region was the single case of 8 successes in 8 trials based on a conventional probability criterion ($< 5\%$).
5. Fisher asserted that no alternative hypothesis was (ever) required.

If and only if the 8 trials produced 8 successes was Fisher willing to reject the null hypothesis effectively acknowledging the Lady's ability with $> 98\%$ confidence (but without quantifying her ability).

Testing Whiteness of Residuals using Autocovariance.

If the model is accurately describing the observed data, then the residuals should be white.

- (Residuals): $\epsilon_t(\theta_n) = y_t - \hat{y}_t(\theta_n|t-1)$ for all $t = 1, \ldots, n$.
- (Innovations): $e_t = y_t - \hat{y}_t(\theta_0|t-1)$ for all $t = 1, \ldots, n$.
- (Noise): $D_t = G^{-1}(q^{-1};\theta_0)\left(Y_t - H(q^{-1};\theta_0)u_t\right)$

A way to verify this is to test for the hypothesis.

$$\begin{cases} \mathcal{H}_0 : & \epsilon_t(\theta_n) \text{ is white} \\ \mathcal{H}_1 : & \text{otherwise} \end{cases}$$

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 Z be a random variable which is distributed as $\chi^2(m)$, define χ^2_α for given α as follows

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

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

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

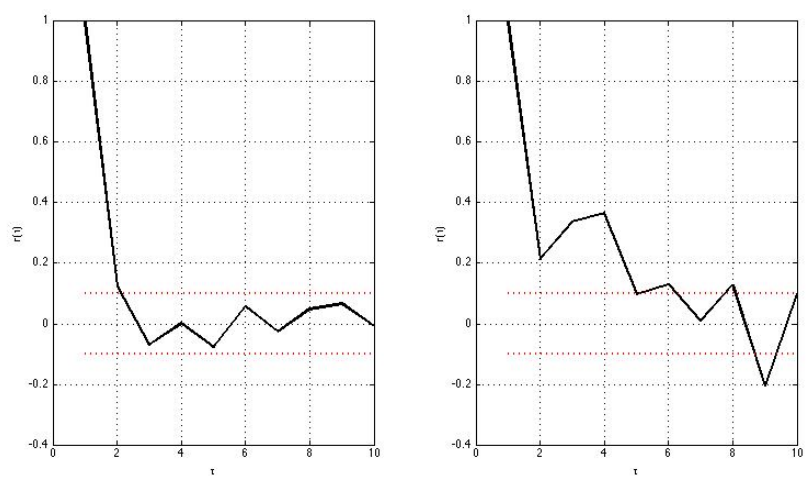


Figure 2: Test of the autocorrelation sequence (a) accept $\mathcal{H}_0$, (b) reject $\mathcal{H}_0$.

Testing Whiteness of Residuals using Zero-Crossings.

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

$$Z_n = \sum_{t=1}^{n-1} I(\epsilon_t \epsilon_{t+1} < 0)$$

then it can be shown that

$$Z_n \to \mathcal{N}(m, p),$$

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

$$\frac{2Z_n - n}{\sqrt{n}} \to \mathcal{N}(0, 1).$$

Testing left-over Correlations of Residuals with Inputs.

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.

Comparing two candidate model structures.

Use the PEM loss function $V_n(\theta_n)$ as a measure of the model quality. For models of increasing 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 d and d'. Then

$$T(Z_n) = n\frac{V_n^2 - V_n'^2}{V_n^2} \to \chi^2(d' - d).$$

Hence we choose model order d at significance level α if

$$T(Z_n) \leq \chi_\alpha^2(d' - d),$$

otherwise select d'.

Information Criteria

Parsimony principle:

'Simple Models are to be preferred.'

Einstein:

'Everything should be made as simple as possible, but not simpler.'

Box:

'All models are wrong, but some are useful'

- If $\theta_n = \theta_0$, then residuals = noise.
- In the least square case, we had that $\mathbb{E}[V_n(\theta_n)] = \frac{n-d}{n}\lambda^2$, so results are biased and residuals $\neq$ noise.

Another approach is to formulate a criterion that is a function of the loss $V_n(\theta_n)$, and which also penalizes the model order:

$$\mathrm{IC}(\mathcal{M}) = V_n(\theta_n)(1 + \beta(n, d))$$

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

Important examples of penalization functions are

1. (Akaike AIC)

$$\mathrm{AIC}(\mathcal{M}) \propto \log V_n(\theta_n) + \frac{2d}{n}$$

2. (Final Prediction Error FPE)

$$\mathrm{FPE}(\mathcal{M}) = V_n(\theta_n) \left(\frac{1 + d/n}{1 - d/n} \right)$$

3. (Minimum Description Length, MDL)

$$\mathrm{MDL}(\mathcal{M}) = V_n(\theta_n)\left(1 + \frac{d\log(n)}{n}\right)$$

The AIC and the FPE are asymptotically equivalent, but it can be shown that both will tend too 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).
- The 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.