Lecture 3

Model Structure Determination and Model Validation –

“A model is of no use if its validity is not verified.”

Choice of Model Structure

1. Type of model set.
   - Ex: Linear or nonlinear model, black-box or white-box model.
   - In our case: ARX, ARMAX, OE, ...
2. Size of the model set. Orders of the polynomials ($A(q^{-1})$, $B(q^{-1})$, $C(q^{-1})$, etc). No true orders in the reality!
3. Model parametrization
   - Transformations of data.
   - Choice of operators: E.g. $q \longleftrightarrow \delta = \frac{q-1}{h}$.

Objective: Obtain a good model at a low cost!

- Quality of model: A scalar measure of the goodness, e.g., the mean-square error (MSE).
  - MSE consists of a bias contribution and a variance contribution.
  - Reduce bias ⇒ more flexible model structures. Decrease variance ⇒ decrease the number of estimated parameters.
  - Trade-off between: Flexibility and parsimony (too complex).
- Price of model:
  - Algorithm complexity.
  - Computational time and power.
- Intended use of the model!

Model Validation

Reasons that it is important to validate the model structure are that

- An underparameterized model is inaccurate/not flexible enough.
- An overparameterized model is not parsimonous and leads to unnecessary complicated computations.

Basic approaches:

- Plots of signals.
- Common sense (a priori information, will the model serve its purpose?)
- Statistical tests.

Basic Plots and Common Sense

- Compare the measured output with the model output:

  $$y_m(t) = G(q^{-1}; \hat{\theta}_N)u(t)$$

  the difference is due to modeling errors and disturbances.
- Plot the difference $\varepsilon(t) = y(t) - y_m(t)$.
- Compare a step response to the modeled step response.
- Compare the estimated transfer function to the transfer function of the model (frequency domain).

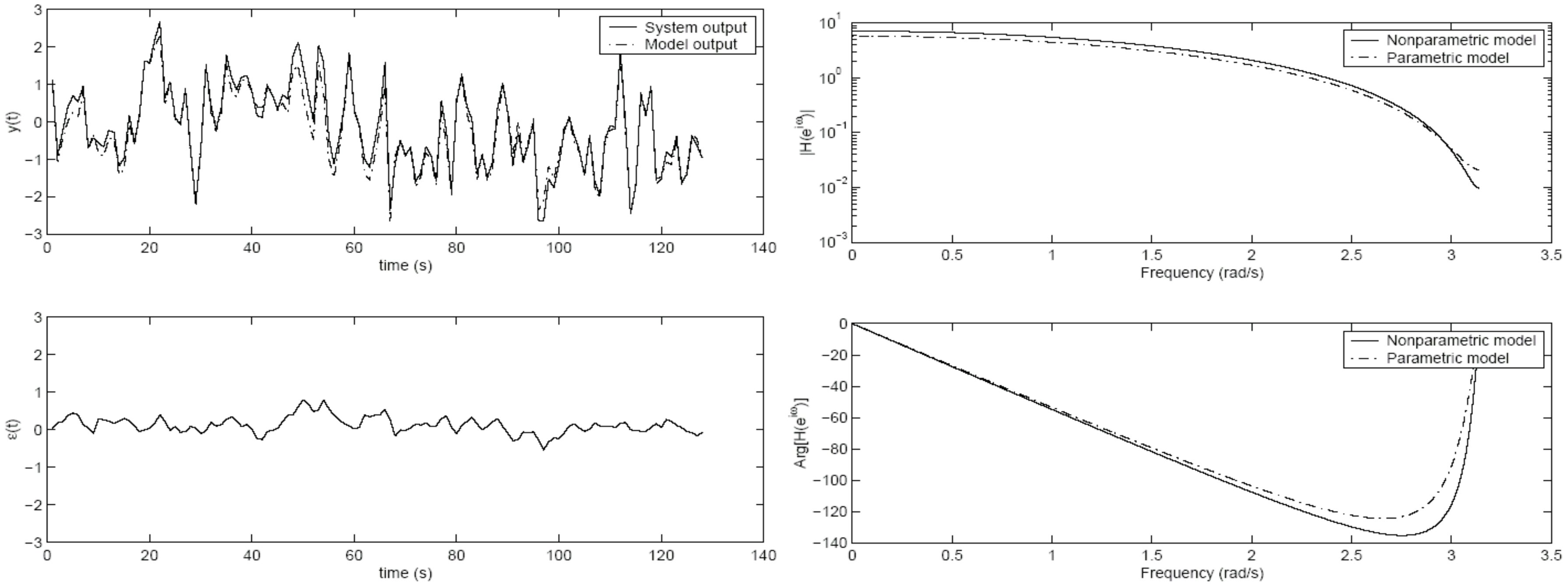

Figure 1: Left: Model output comparison. Right: Transfer function comparison.

What to compare?

Def: The k-step ahead model predictions $\hat{y}_k(t;\hat{\boldsymbol{\theta}})$ are based on the past data

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

using the model associated with $\hat{\boldsymbol{\theta}}$.

Rem: Common choices are:

- $\hat{y}_1(t;\hat{\boldsymbol{\theta}})$ is the standard mean square optimal predictor

  $$\hat{y}_1(t;\hat{\boldsymbol{\theta}}) = \hat{y}(t|t-1,\hat{\boldsymbol{\theta}}) = \hat{H}^{-1}(q^{-1})\hat{G}(q^{-1})u(t)+\big(1-\hat{H}^{-1}(q^{-1})\big)y(t)$$

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

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

To compare different models we often use a scalar measure of the differences $y(t) - \hat{y}_k(t|\hat{\boldsymbol{\theta}}_N)$

$$V_k(\hat{\boldsymbol{\theta}}_N) = \frac{1}{N}\sum_{t=1}^{N} |y(t) - \hat{y}_k(t|\hat{\boldsymbol{\theta}}_N)|^2$$

Example: What are the properties of

$$\hat{y}(t) = y(t-1)$$

Questions to Answer

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

- Is the model flexible enough? Is the model structure large enough to cover the true system?
- Is a given model too complex?
- Which model structure of two candidates should be chosen?

Each of these questions have several different answers; no solution is perfect.

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

$$\varepsilon(t) = \varepsilon(t, \hat{\boldsymbol{\theta}}_N) = y(t) - \hat{y}(t|t-1, \hat{\boldsymbol{\theta}}_N)$$

Rem: The residuals are the prediction errors evaluated at $\hat{\boldsymbol{\theta}}_N$. If $\hat{\boldsymbol{\theta}}_N = \boldsymbol{\theta}_0$ then $\varepsilon(t)$ is white!

- If

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

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

- The covariance between residuals and past inputs

$$\hat{R}_{\varepsilon u}(\tau) = \frac{1}{N}\sum_{t=\tau}^{N} \varepsilon(t)u(t-\tau)$$

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

Testing Whiteness

If the model is accurately describing the observed data, then the residuals $\varepsilon(t)$ should be white. A way to validate the model is thus to, in some way, test the hypotheses

$$\begin{aligned} H_0 \quad &: \quad \varepsilon(t) \text{ is a white sequence} \\ H_1 \quad &: \quad \varepsilon(t) \text{ is not a white sequence} \end{aligned}$$

This can be done in several ways, for example:

Autocorrelation Test

The autocovariance of the residuals is estimated as:

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

If H_0 holds, then the squared covariance estimates is asymptotically χ^2 distributed:

$$\frac{N}{\hat{r}_\varepsilon^2(0)}\sum_{i=1}^{m} \hat{r}_\varepsilon^2(i) \rightarrow \chi^2(m)$$

Furthermore, the normalized autocovariance estimates are asymptotically Gaussian distributed

$$\sqrt{N}\frac{\hat{r}_\varepsilon(\tau)}{\hat{r}_\varepsilon(0)} \rightarrow N(0,1)$$

A typical way of using the first test statistics for validation is as follows (the second can be used similarly).

Let x denote a random variable which is χ^2-distributed with m degrees of freedom. Furthermore, define $\chi^2_\alpha(m)$ by

$$\alpha = P(x > \chi^2_\alpha(m))$$

for some given α (typically between 0.01 and 0.1). Then if,

$$\frac{N}{\hat{r}^2_\varepsilon(0)} \sum_{i=1}^{m} \hat{r}^2_\varepsilon(i) > \chi^2_\alpha(m) \quad \text{reject } H_0$$

$$\frac{N}{\hat{r}^2_\varepsilon(0)} \sum_{i=1}^{m} \hat{r}^2_\varepsilon(i) \leq \chi^2_\alpha(m) \quad \text{accept } H_0$$

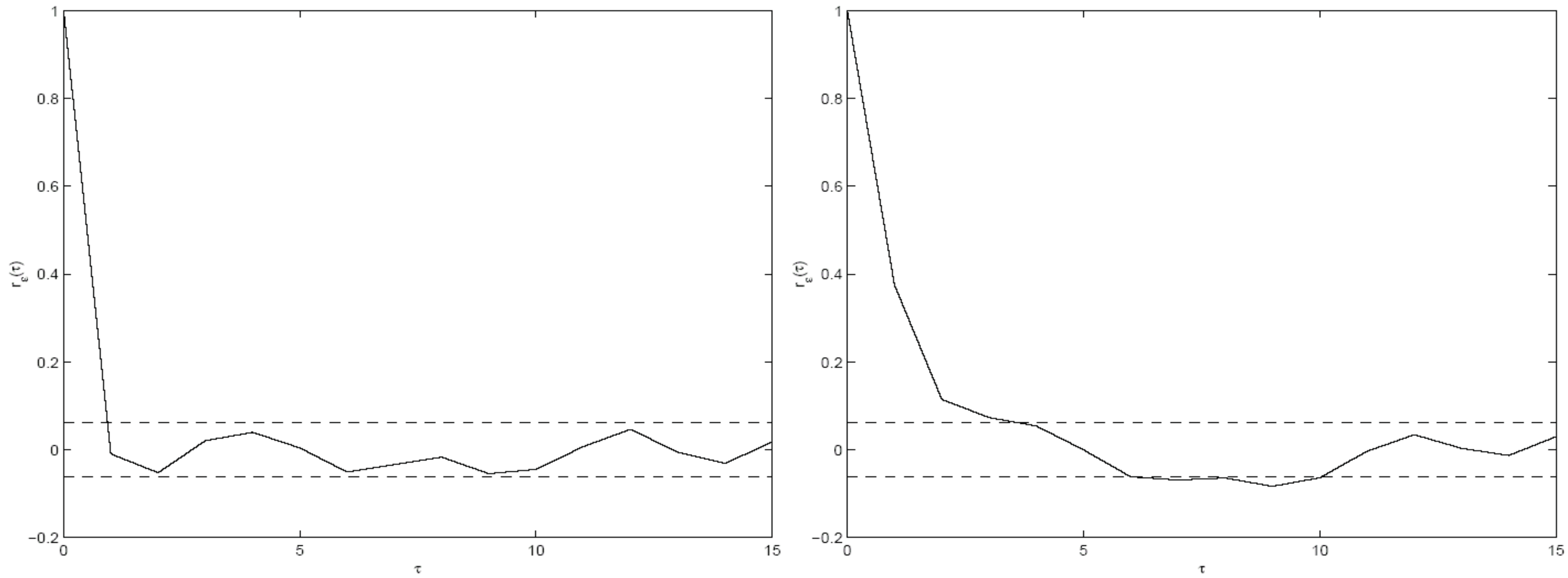

Figure 2: Autocorrelation test. Left: White residuals. Right: Correlated residuals.

Cross Correlation Test

If the model is an accurate description of the system, then the input and the residuals should be uncorrelated (no unmodeled dynamics), i.e.,

$$r_{\varepsilon u}(\tau) = E\varepsilon(t+\tau)u(t) = 0$$

- If $r_{\varepsilon u}(\tau) \neq 0$ for $\tau < 0$ then there is output feedback in the input.
- Indication of wrong time delay in model. If a time delay of two samples has been assumed in the model, but the true delay is one sample, then a clear correlation between $u(t-1)$ and $\varepsilon(t)$ will show up.
- To visualize the correlation, it might be better to postulate a model like $\varepsilon(t) = G_\varepsilon(q)u(t)$.

The following results can be used to design a hypothesis test whether the input and the residuals are uncorrelated.

Form the normalized test quantity

$$x_\tau = \frac{\hat{r}^2_{\varepsilon u}(\tau)}{\hat{r}_\varepsilon(0)\hat{r}_u(0)}$$

where $\hat{r}_{\varepsilon u}(\tau)$ is the estimated crosscovariance

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

Is a model too complex?

It is important to detect if a model is overparameterized as such a model is unnecessarily complicated and can be sensitive to parameter variations. One way to do so is to study a pole-zero plot of the model transfer function.

If there are signs of pole-zero cancellation for model orders higher than a certain threshold n, it suggests that $p \leq n$ is a suitable model order for the system.

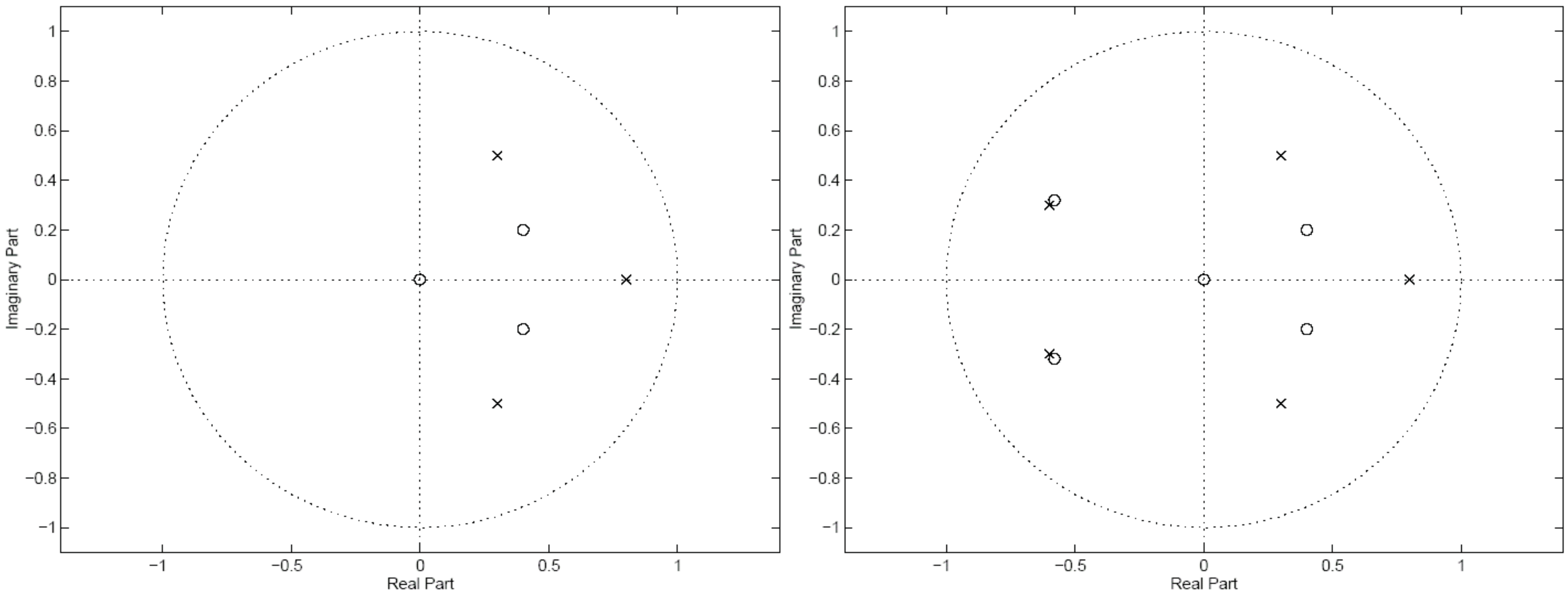

Figure 3: Pole-Zero cancellation. Left: ARX(3,2). Right: ARX(5,4)

Cross validation

- Check the criterion $V_1(\hat{\boldsymbol{\theta}}_N)$ (or $V_k(\hat{\boldsymbol{\theta}}_N)$). A model structure that is "too rich" to describe the system will also partly model the disturbances that are present in the actual data set. This is called an "overfit" of the data.
- Using a fresh dataset that was not included in the identification experiment for model validation is called "cross validation".
- Cross validation is a nice and simple way to compare models and to detect "overfitted" models.
- Cross validation requires a "large amount" of data, the validation data cannot be used in the identification.

The parsimony principle

The parsimony principle states that one should not use extra parameters to model a system if they are not necessary.

Assume that the quality of the model is measured by W_N:

$$W_N(\hat{\boldsymbol{\theta}}_N) = E\varepsilon^2(t, \hat{\theta}_N)$$

where $\varepsilon(t, \theta)$ is the prediction error. If the estimate is exact $\hat{\boldsymbol{\theta}}_N = \boldsymbol{\theta}_0$, the prediction error would be white, and $W_N(\hat{\boldsymbol{\theta}}_N) = \lambda_0^2$. However, $\hat{\boldsymbol{\theta}}_N$ deviates somewhat from $\boldsymbol{\theta}_0$.

Results: It holds that

$$E\,W_N(\hat{\boldsymbol{\theta}}_N) \approx V_1(\hat{\boldsymbol{\theta}}_N) + \lambda_0^2 \frac{2p}{N}$$

$$V_1(\hat{\boldsymbol{\theta}}_N) \approx \lambda_0^2\big(1 - \frac{p}{N}\big)$$

where λ_0^2 is the variance of the disturbance and $p = \dim \boldsymbol{\theta}$.

Rem:

- $E\,W_N(\hat{\boldsymbol{\theta}}_N)$ represents the average as the estimated models are evaluated on validation data.
- $V_1(\hat{\boldsymbol{\theta}}_N)$ will decrease with increasing number of parameters. However, each parameter carries a variance penalty that will contribute with $2\lambda_0^2/N$.

Another approach is to formulate a criterion that is a function of the loss function $V_1(\boldsymbol{\theta})$, but also penalizes the model order

$$W_N = V_1(\hat{\boldsymbol{\theta}}_N)\big[1 + \beta(N,p)\big]$$

where $\beta(N,p)$ is a function which should increase with the model order p (to penalize too complex model structures), but decrease to zero when $N \to \infty$.

Important examples of penalty functions are:

(i) The Akaike information criterion (AIC)

$$\text{AIC}(p) = V_1(\hat{\boldsymbol{\theta}}_N)\Big[1 + \frac{2p}{N}\Big]$$

(ii) The final prediction error criterion (FPE)

$$\text{FPE}(p) = V_1(\hat{\boldsymbol{\theta}}_N)\Big[\frac{1 + p/N}{1 - p/N}\Big]$$

(iii) The minimum description length (MDL)

$$\mathrm{MDL}(p) = V_1(\hat{\boldsymbol{\theta}}_N)\left[1 + \frac{p \ln N}{N}\right]$$

The AIC and FPE are asymptotically equivalent, but it can be shown that both will tend to choose too high model orders (the estimates are not consistent). The MDL yields estimates that are consistent.

Physical insight might significantly simplify the model order selection.

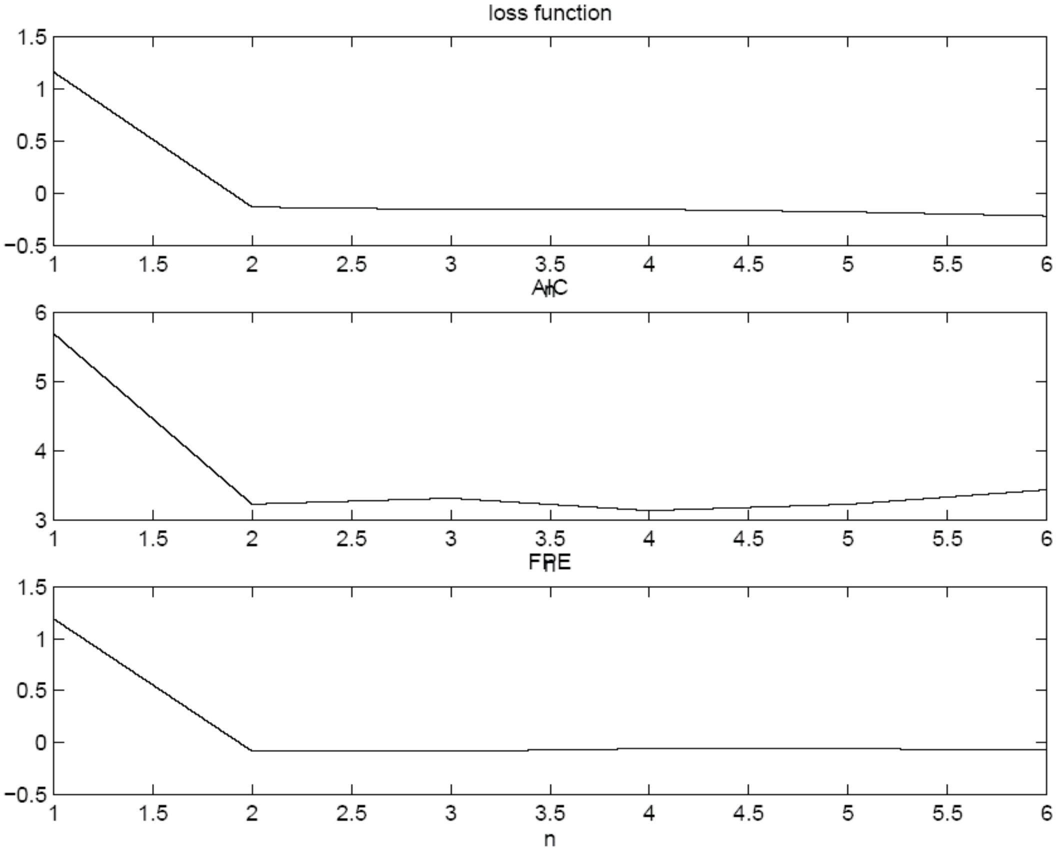

Figure 4: Model structure determination

Summary - Model Parameterizations

- Many different tests can be performed to verify the validity of a model (try simple things first).
- The choice of the appropriate model structure (model order) can be based on statistical tests on the residuals (autocovariance test/ cross-covariance test).
- To decide the appropriate model order tests such as the AIC, FPE or MDL criteria can be used.
- Cross validation is a good approach that should be used if there is a sufficient amount of data available.
- Most tests are implemented in the system identification toolbox for MATLAB.

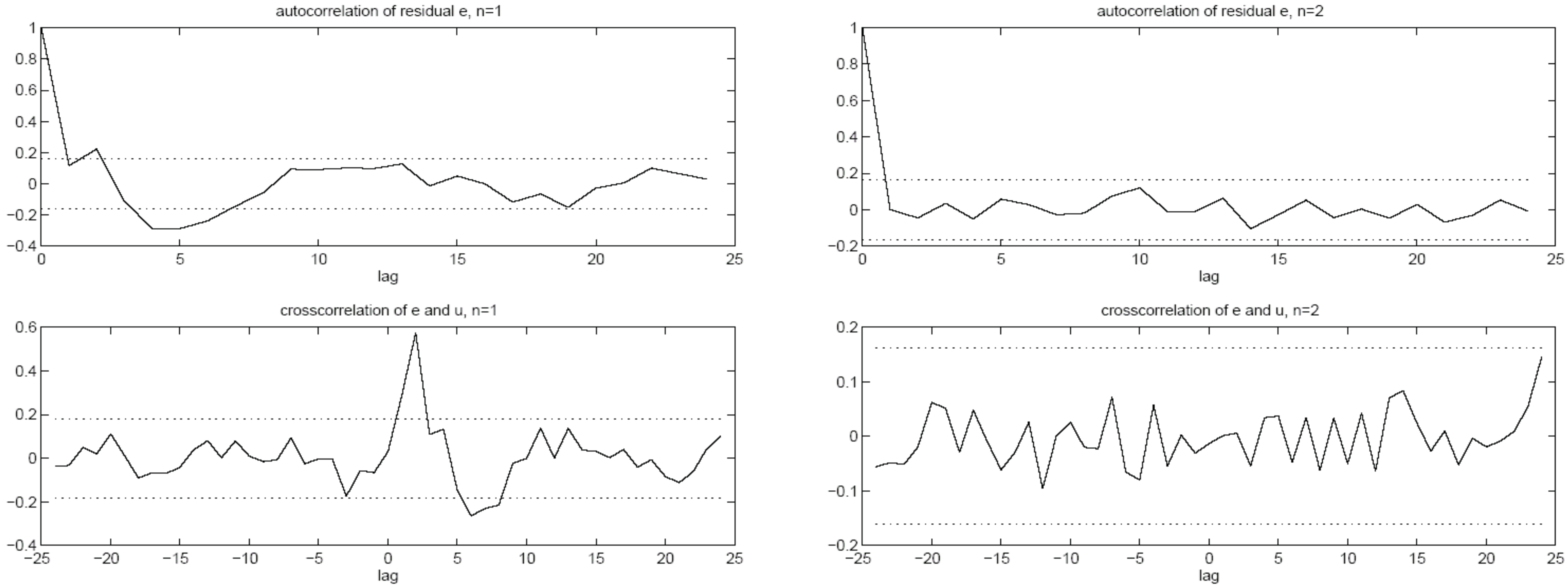

Figure 5: Correlation tests.

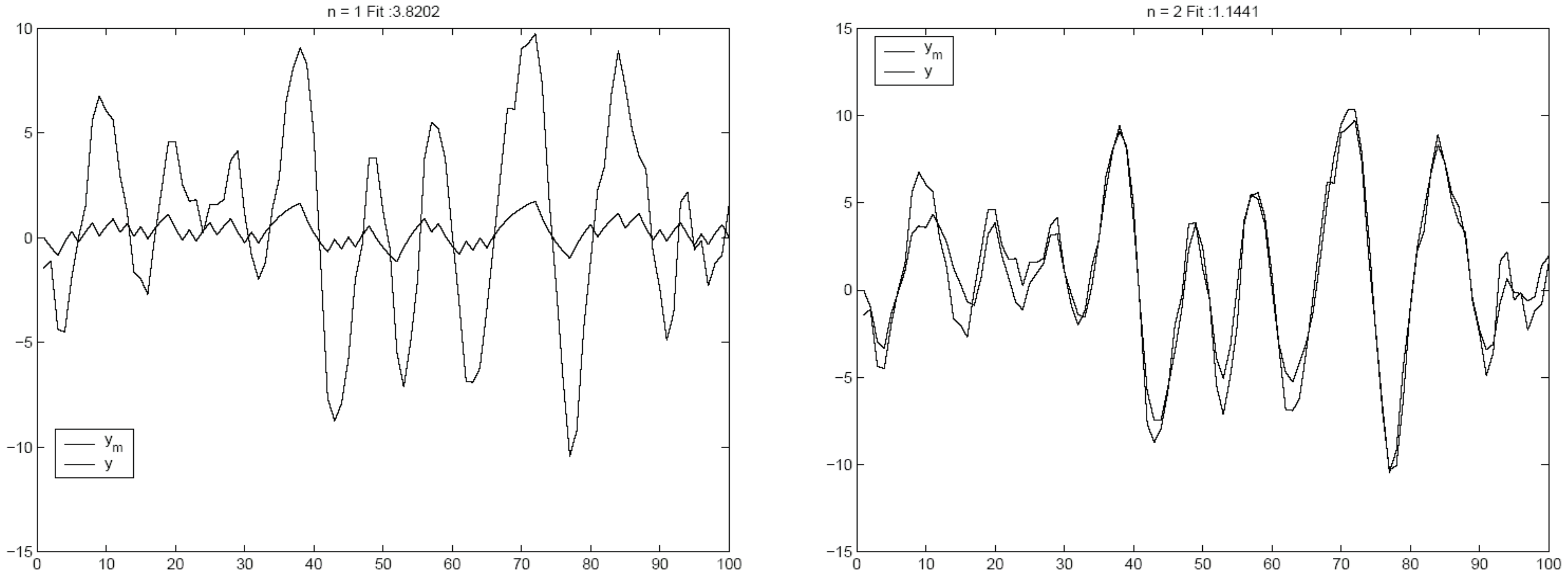

Figure 6: Comparing outputs. Estimation data.

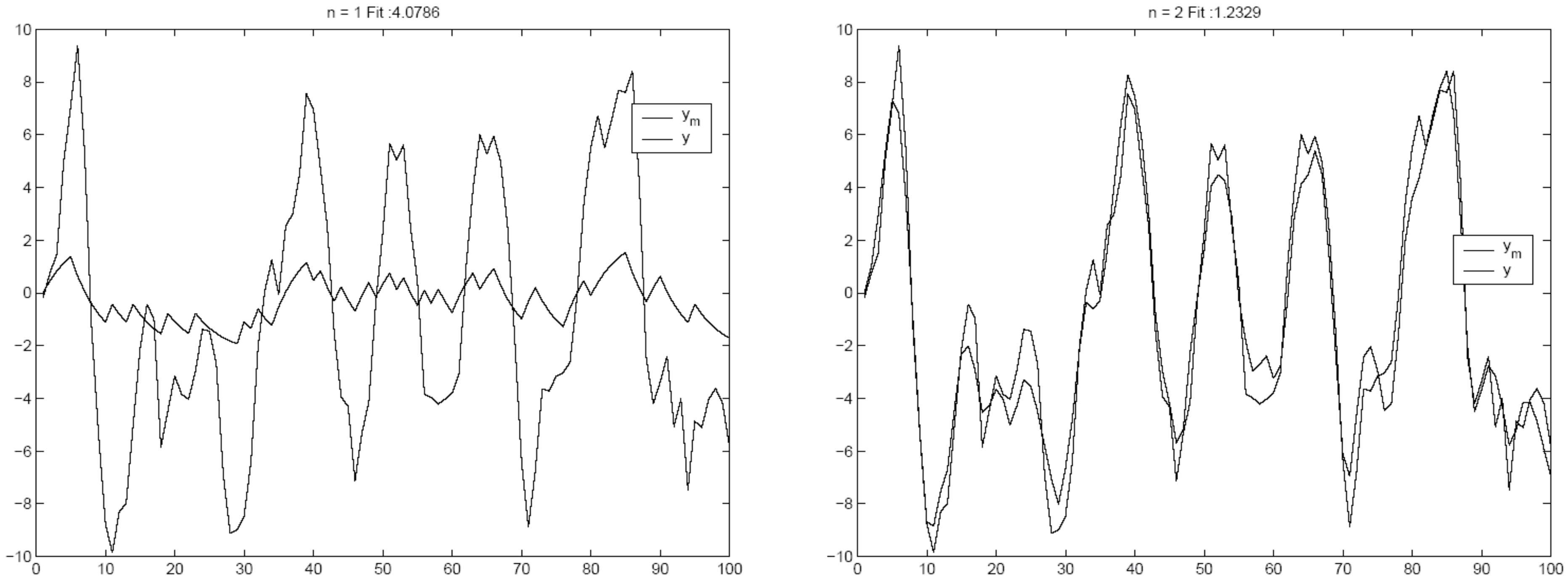

Figure 7: Comparing outputs. Validation data.

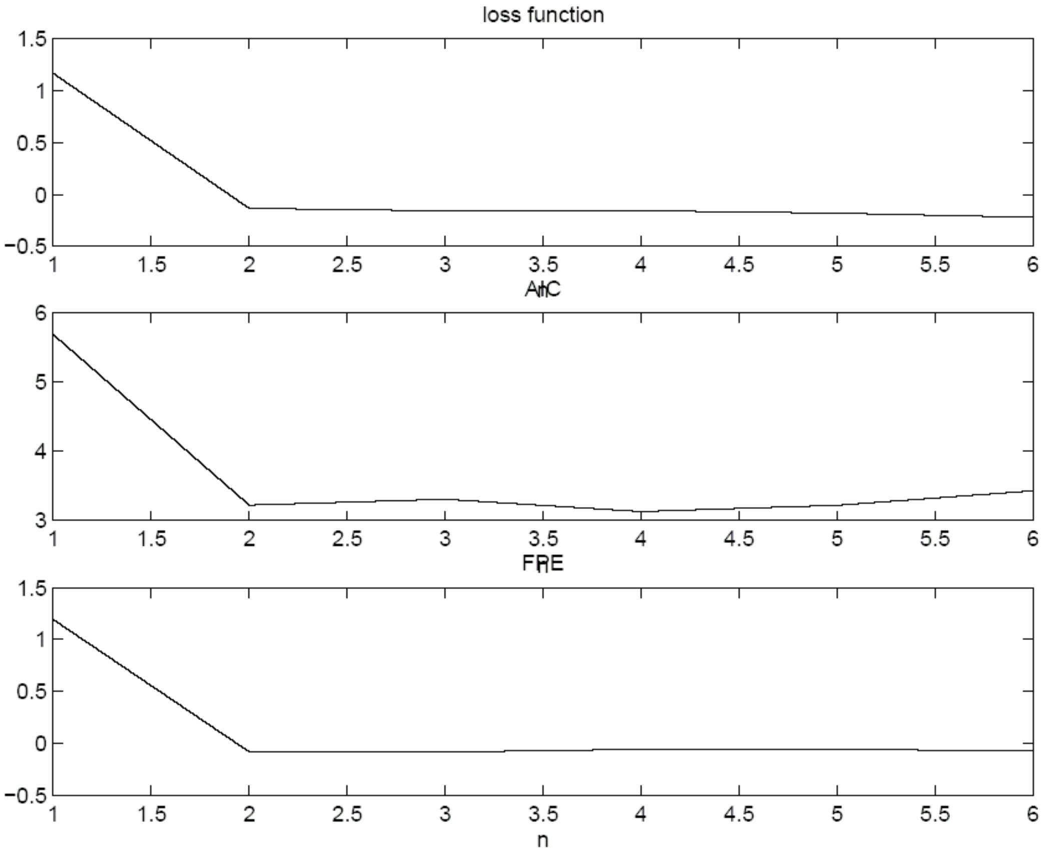

Figure 8: Loss functions.

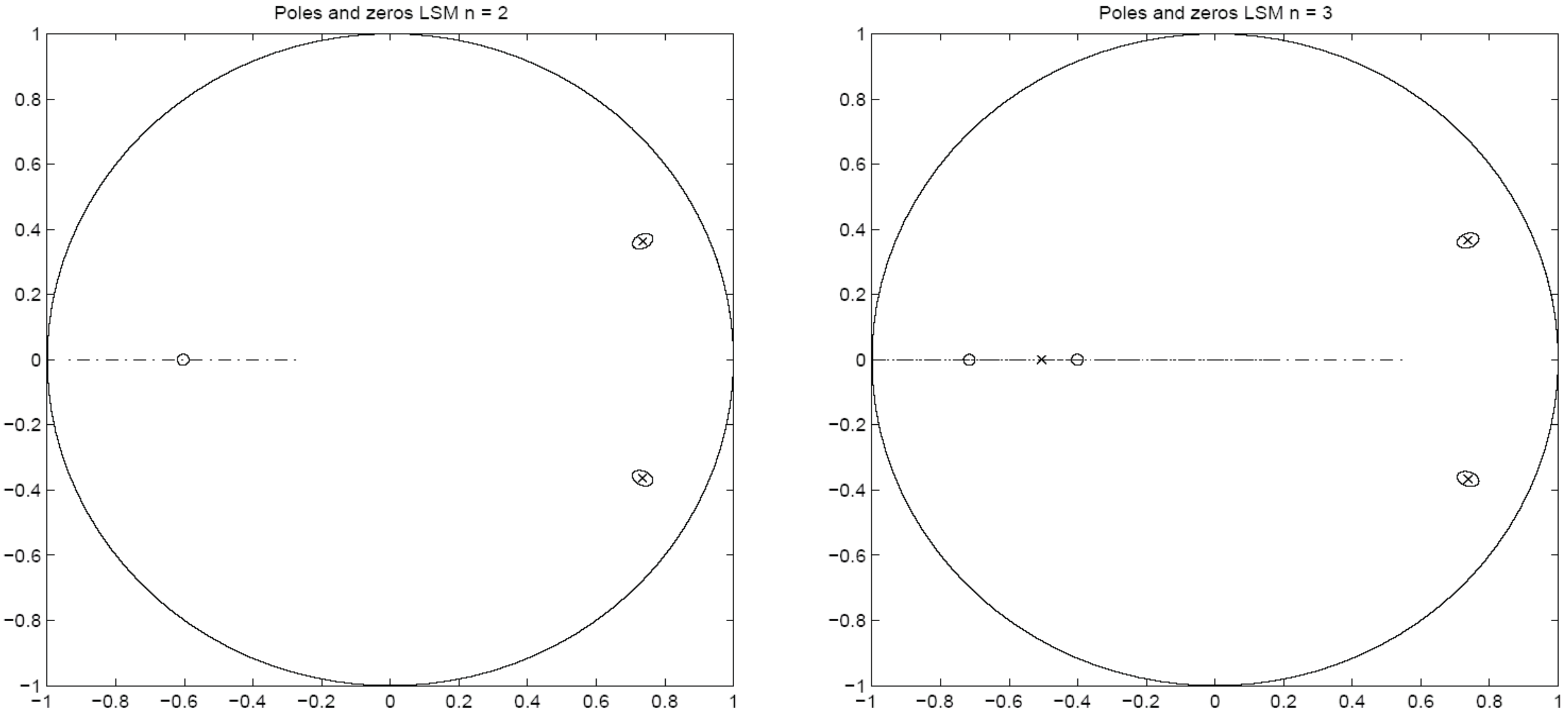

Figure 9: Poles-zeros plots.