14. Model Selection and Model Validation

- Introduction
- Model selection
- Model validation

General aspects of the choice of model structure

1. Type of model set
   - Linear/Nonlinear, State-spaces, black-box models
   - ARX, ARMAX, OE,...
2. Size of the model set
   - the degrees of polynomials $A(q^{-1}), B(q^{-1}), C(q^{-1})$, ...
   - the dimension of state-space models
3. Model parametrization

Objective: Obtain a good model at a low cost

1. **Quality of the model:** defined by a measure of the goodness, e.g., the mean-squared error
   - MSE consists of a *bias* and a *variance* contribution
   - To reduce the bias, one has to use more flexible model structures (requiring more parameters)
   - The variance typically increases with the number of estimated parameters
   - The best model structure is therefore a trade-off between *flexibility* and *parsimony*
2. **Price of the model:** an estimation method (which typically results in an optimization problem) highly depends on the model structures, which influences:
   - Algorithm complexity
   - Properties of the loss function
3. Intended use of the model

The Bias-Variance decomposition

Assume that the observation Y obeys

$$Y = f(X) + \nu, \quad \mathbf{E}\,\nu = 0, \quad \mathbf{cov}(\nu) = \sigma^2$$

The mean-squared error of a regression fit $\hat{f}(X)$ at $X = x_0$ is

$$\begin{aligned}\mathsf{MSE} &= \mathbf{E}[(Y - \hat{f}(x_0))^2 | X = x_0] \\ &= \sigma^2 + [\mathbf{E}\,\hat{f}(x_0) - f(x_0)]^2 + \mathbf{E}[\hat{f}(x_0) - \mathbf{E}\,\hat{f}(x_0)]^2 \\ &= \sigma^2 + \mathsf{Bias}^2(\hat{f}(x_0)) + \mathsf{Var}(\hat{f}(x_0))\end{aligned}$$

- This relation is known as *Bias-Variance decomposition*
- No matter how well we estimate $f(x_0)$, σ^2 represents *irreducible error*
- Typically, the more complex we make model $\hat{f}$, the lower the bias, but the higher the variance

Example

Consider a stable first-order AR process

$$y(t) + ay(t-1) = \nu(t)$$

where $\nu(t)$ is white noise with zero mean and variance λ^2

Consider the following two models:

$$\begin{aligned}\mathcal{M}_1 \quad &: \quad y(t) + a_1 y(t-1) = e(t) \\ \mathcal{M}_2 \quad &: \quad y(t) + c_1 y(t-1) + c_2 y(t-2) = e(t)\end{aligned}$$

Let $\hat{a}_1, \hat{c}_1, \hat{c}_2$ be the LS estimates of each model

We can show that

$$\mathrm{Var}(a_1) < \mathrm{Var}(c_1)$$

(the simpler model has less variance)

Apply a linear regression to the dynamical models

$$y(t) = H(t)\theta + \nu(t)$$

It asymptotically holds that

$$\text{Cov}(\hat{\theta}) = \lambda^2 [\mathbf{E}\, H(t)^* H(t)]^{-1}$$

For model $\mathcal{M}_1$, we have $H(t) = -y(t-1)$, so

$$\text{Cov}(\hat{a}_1) = \lambda^2 / R_y(0)$$

For model $\mathcal{M}_2$, we have $H(t) = -\begin{bmatrix} y(t-1) & y(t-2) \end{bmatrix}$ and

$$\text{Cov}\begin{pmatrix} \hat{c}_1 \\ \hat{c}_2 \end{pmatrix} = \lambda^2 \begin{bmatrix} R_y(0) & R_y(1) \\ R_y(1) & R_y(0) \end{bmatrix}^{-1}$$

To compute $R_y(\tau)$, we use the relationship

$$R_y(\tau) = (-a)^\tau R_y(0),$$

where $R_y(0)$ is solved from a Riccati equation and the solution is

$$R_y(0) = \frac{\lambda^2}{1-a^2}$$

Apply this result, we can show that

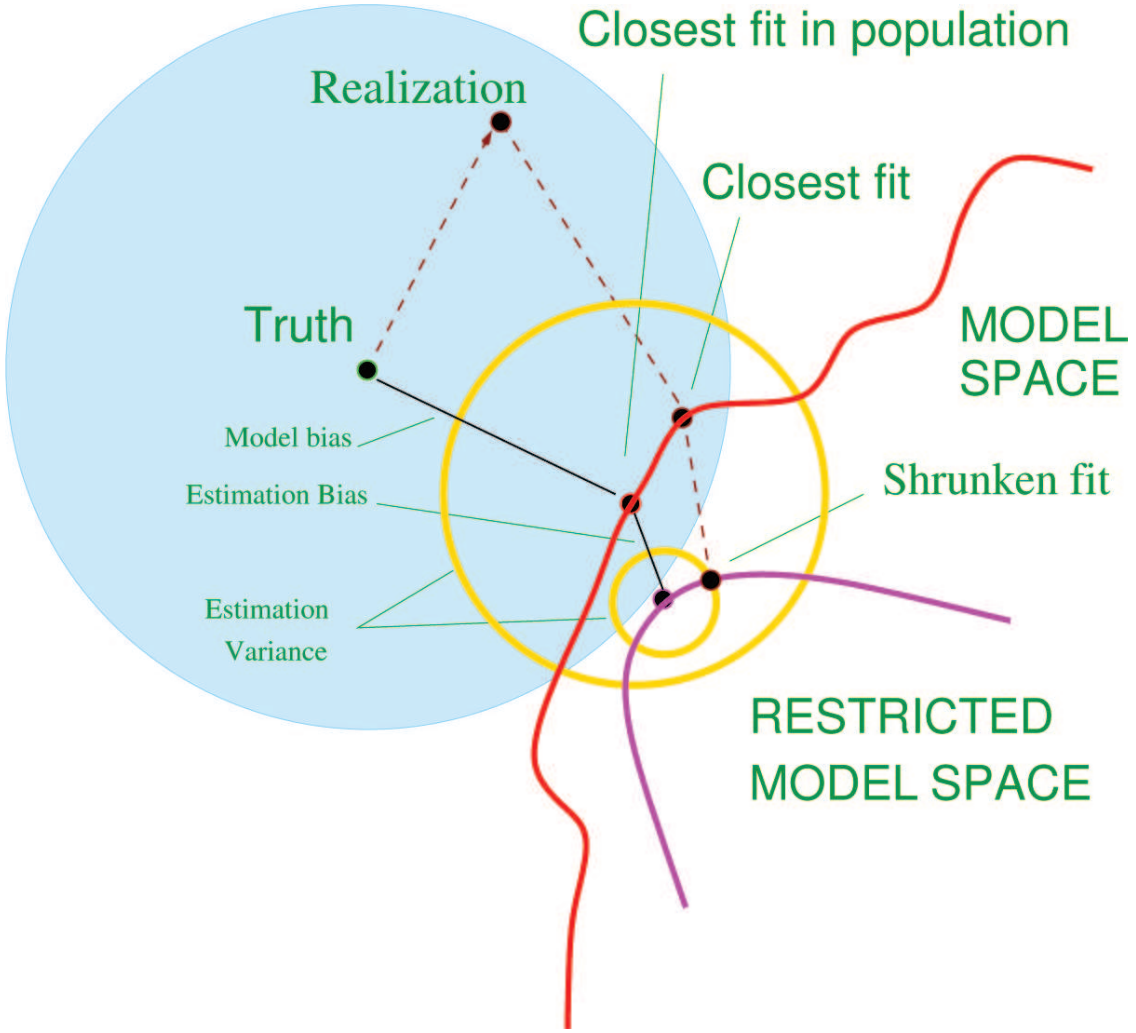
$$\mathsf{Cov}(\hat{c}_1) = \frac{\lambda^2 R_y(0)}{R_y(0)^2 - R_y(1)^2} = \frac{\lambda^2}{R_y(0)(1-a^2)}$$

while

$$\mathsf{Cov}(\hat{a}_1) = \frac{\lambda^2}{R_y(0)}$$

Since $|a| < 1$, we can claim that $\mathsf{Cov}(\hat{a}_1) < \mathsf{Cov}(\hat{c}_1)$

Schematic of the behavior of bias and variance



(T. Hastie *et.al.* *The Elements of Statistical Learning*, Springer, 2010 page 225)

Model selection

- Simple approach: enumerate a number of different models and to compare the resulting models
- What to compare ? how well the model is capable of reproducing these data
- How to compare ? comparing models on fresh data set: cross-validation
- Model selection criterions
  - Akaike Information Criterion (AIC)
  - Baysian Information Criterion (BIC)
  - Minimum Description Length (MDL)

Overfitting

We start by an example of AR model with white noise ν

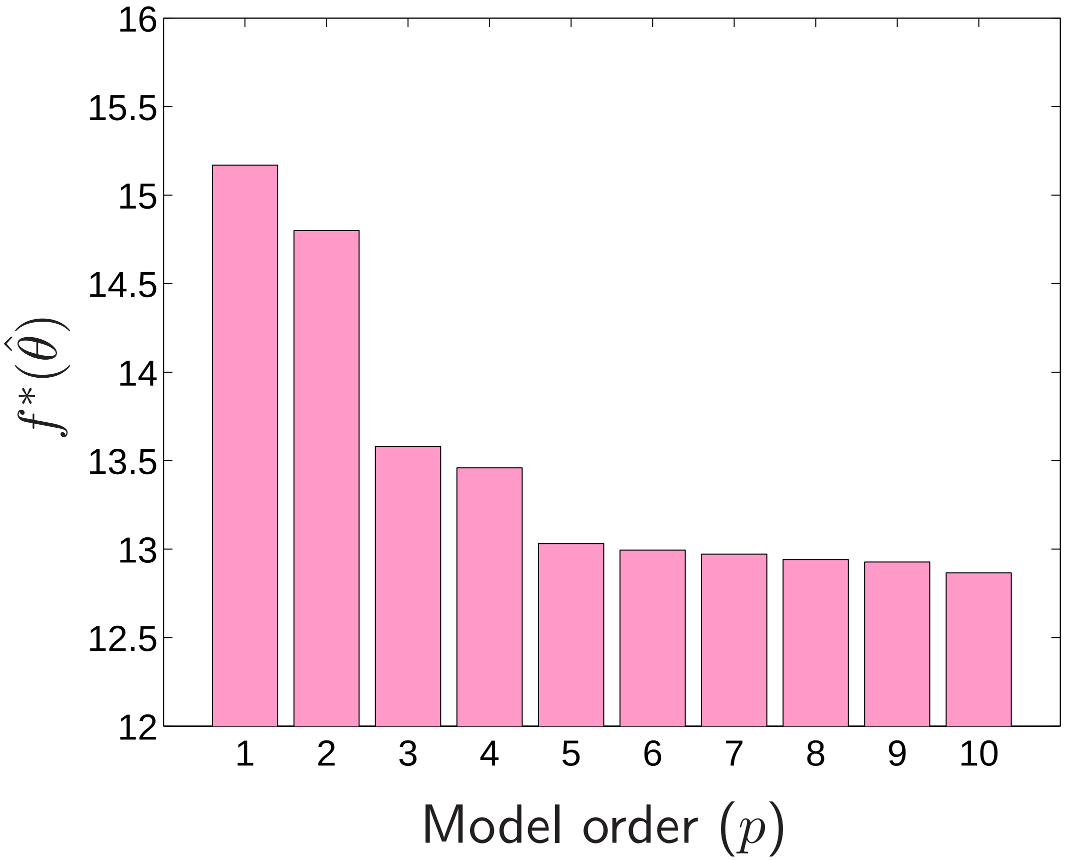
$$y(t) + a_1 y(t-1) + \dots a_p y(t-p) = \nu(t)$$

- The true AR model has order $p = 5$
- The parameters to be estimated are $\theta = (a_1, a_2, \dots, a_p)$ with p unknown
- Question: How to choose a proper value of p ?
- Define a quadratic loss function

$$f(\theta) = \sum_{t=p+1}^{N} |y(t) - (a_1 y(t-1) + \dots + a_p y(t-p))|^2$$

and obtain $\hat{\theta}$ by using the LS method:

$$\hat{\theta} = (\hat{a}_1, \hat{a}_2, \dots, \hat{a}_p) = \underset{\theta}{\operatorname{argmin}}\, f(\theta)$$



- The minimized loss is a decreasing function of the model structure
- f begins to decreases as the model picks up the relevant features
- As p increases, the model tends to *over fit* the data
- In practice, we look for the “knee” in the curve (around $p = 5$)

The Parsimony Principle

Among competing models which all explain the data well, the model with the smallest number of parameters should be chosen

In the previous example, how to determine the model order p ?

- a trade-off curve between the loss function and the model order
- model selection criterions

A model selction criterion consists of two parts:

$$\text{Loss function} + \text{Model complexity}$$

- The first term is to assess the quality of the model, e.g., the quadratic loss, the likelihood function
- The second term is to penalize the model order and grows as the number of parameters increases

Examples of model selection criterions

Akaike Information Criterion (AIC)

$$\mathsf{AIC} = -2\mathcal{L} + 2d$$

Bayesian Information Criterion (BIC)

$$\mathsf{BIC} = -2\mathcal{L} + d\log N$$

Akaike's Final Prediction-Error Criterion (FPE)

$$\mathsf{FPE} = \left(\frac{1}{N}\sum_{t=1}^{N} e^2(t,\theta)\right)\frac{1+d/N}{1-d/N}$$

- $\mathcal{L}$ is the loglikelihood function
- d is the number of effective parameters
- $e(t,\theta)$ is the prediction error

Some known properties:

- BIC tends to penalize complex models more heavily (due to the term $\log N$)

- BIC is asymptotically consistent

  (the probability that BIC will select the correct model approaches one as the sample size $N \to \infty$)

- On the other hand, AIC and FPE tends to choose models which are too complex as $N \to \infty$

AIC and BIC for Gaussian innovation

As we have seen in PEM (page 11-9 to 11-10), the ML method can be interpreted as PEM if the noise is *Gaussian*

In this case, the loglikelihood function (up to a constant) is

$$\mathcal{L} = \log L(\theta) = -\frac{N}{2} \log \det R(\theta)$$

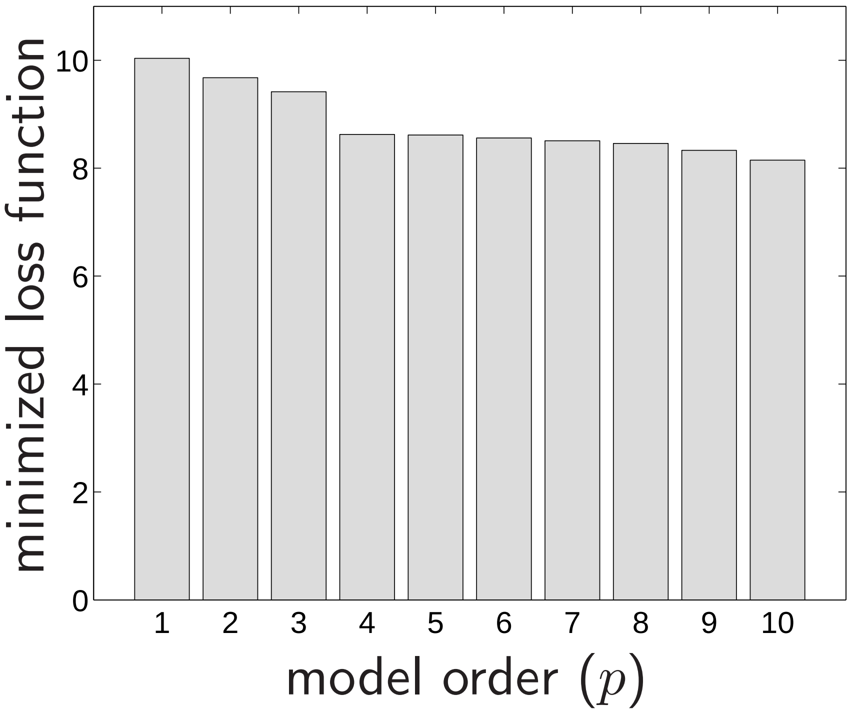
where $R(\theta) = \frac{1}{N} \sum_{t=1}^{N} e(t,\theta)e(t,\theta)^*$ is the sample covariance matrix

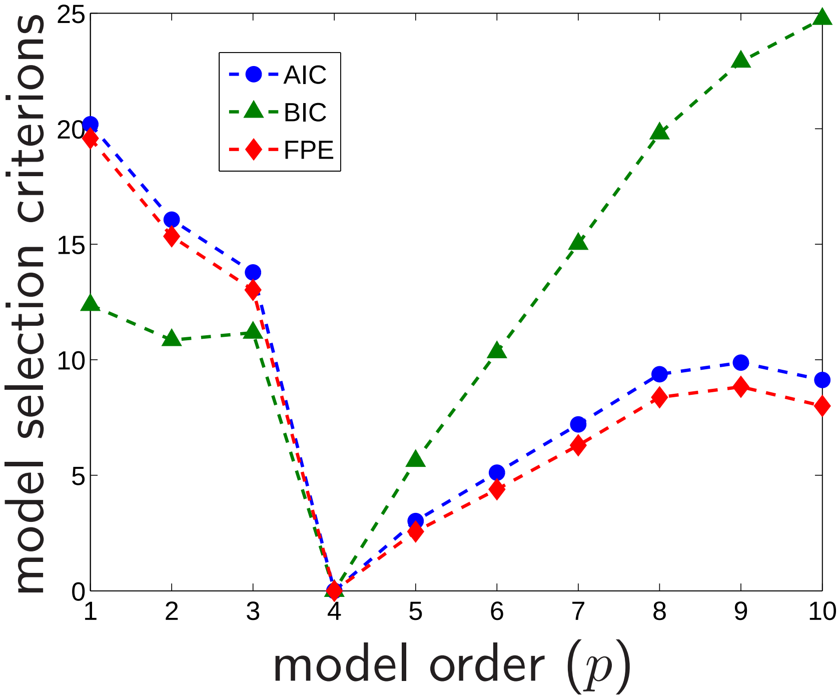
For scalar case, substituting $\mathcal{L}$ in AIC and BIC gives

$$\text{AIC} = -2\mathcal{L} + 2d = N \log \left(\frac{1}{N} \sum_{t=1}^{N} e^2(t,\theta) \right) + 2d$$

$$\text{BIC} = -2\mathcal{L} + d \log N = N \log \left(\frac{1}{N} \sum_{t=1}^{N} e^2(t,\theta) \right) + d \log N$$

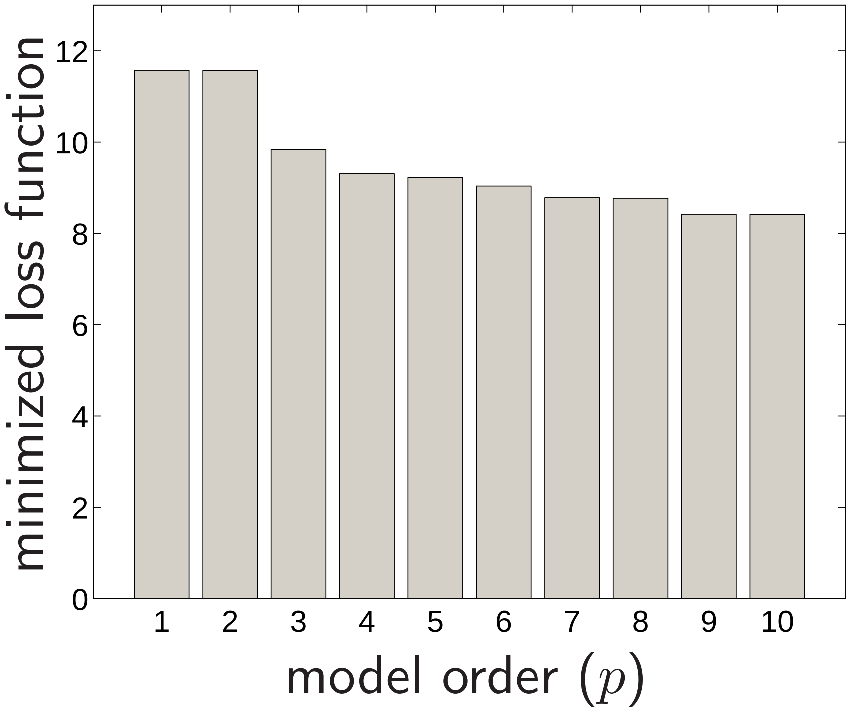
Example

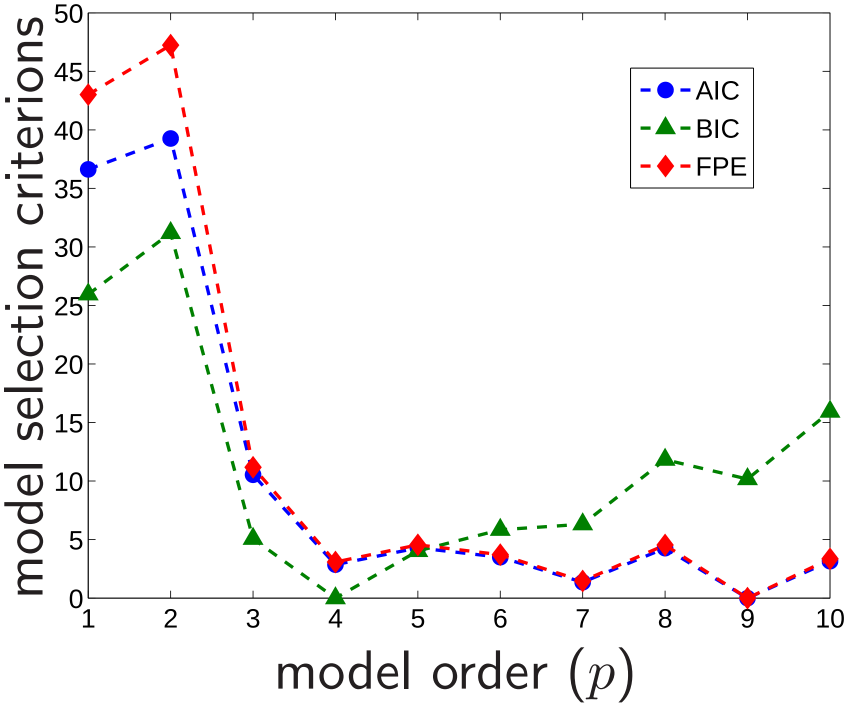




- The true system is AR model of order 4 with white noise of variance 1
- Generate data of 100 points and estimate θ using LS

Another realization





- AIC an FPE pick model of order 9 (too high)
- BIC still choose the correct model order (4)
- The estimates from AIC and FPE are not consistent
- BIC yields estimates that are consistent

Model validation

The parameter estimation procedure picks out the *best* model

The problem of model validation is to verify whether *this best* model is "good enough"

General aspects of model validation

- Validation with respect to the purpose of the modeling
- Feasibility of physical parameters
- Consistency of model input-output behavior
- Model reduction
- Parameter confidence intervals
- Simulation and prediction

Comparing model structures

We generally use k*-step ahead model predictions* as the basis of the comparisons

$\hat{y}_k(t|m)$ denotes the k-step predictor based on model m and

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

For a linear model $y = \hat{G}u + \hat{H}\nu$, common choices are

- $\hat{y}_1(t|m)$ is the standard mean square optimal predictor

$$\hat{y}_1(t|m) = \hat{y}(t|t-1) = \hat{H}^{-1}(q^{-1})\hat{G}(q^{-1})u(t) + (1 - \hat{H}^{-1}(q^{-1}))y(t)$$

- $\hat{y}_\infty(t|m)$ is based on past inputs only (referred to as a pure *simulation*)

$$\hat{y}_\infty(t|m) = \hat{G}(q^{-1})u(t)$$

Now the models can be compared via a scalar measure of goodness:

$$J(m) = \frac{1}{N} \sum_{t=1}^{N} \|y(t) - \hat{y}_k(t|m)\|^2$$

The normalized measure, R, is given by detrending y and computing

$$R^2 = 1 - \frac{J(m)}{(1/N) \sum_{t=1}^{N} \|y(t)\|^2}$$

R represents part of the output variation that is explained by the model

- $J(m)$ depends on the realization of the data used in the comparison
- It is natural to consider the expected value of this measure:

$$\bar{J}(m) = \mathbf{E}\, J(m)$$

which gives a quality measure for the given model

Cross validation

- 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

K-fold cross-validation

- A simple and widely used method for estimating prediction error
- Used when data are often scarce, then we split the data into K equal-sized parts
- For the kth part, we fit the model to the other $K-1$ parts of the data
- Then compute $J(m)$ on the kth part of the data
- Repeat this step for $k = 1, 2, \ldots, K$
- The cross-validation estimate of $J(m)$ is

$$\mathsf{CV}(m) = \frac{1}{K}\sum_{i=1}^{K} J_k(m)$$

- If $K = N$, it is known as *leave-one-out* cross-validation

Residual Analysis

The prediction error evaluated at $\hat{\theta}$ is called *the residuals*

$$e(t) = e(t, \hat{\theta}) = y(t) - \hat{y}(t; \hat{\theta})$$

- represents part of the data that the model could not reproduce
- If $\hat{\theta}$ is the true value, then $e(t)$ is white

A pragmatic view starting point is to use the basis statistics:

$$S_1 = \max_t |e(t)|, \quad S_2 = \frac{1}{N} \sum_{t=1}^{N} e^2(t)$$

to asses the quality of the model

The use of these statistics has an implicit invariance assumption

The residuals do not depend on the particular input

- The covariance between the residuals and past inputs

$$R_{eu}(\tau) = \frac{1}{N}\sum_{t=\tau}^{N} e(t)u(t-\tau)$$

  should be small if the model has picked up the essential part of the dynamics from u to y

- It also indicates that the residual is invariant to various inputs

- If

$$R_{e}(\tau) = \frac{1}{N}\sum_{t=\tau}^{N} e(t)e(t-\tau)$$

  is not small for $\tau \neq 0$, then part of $e(t)$ could have been predicted from past data

- This means $y(t)$ could have been better predicted

Whiteness test

If the model is accurately describing the observed data,

then the residuals $e(t)$ should be *white*, $i.e.$,

its covariance function $R_e(\tau)$ is zero except at $\tau = 0$

A way to validate the model is to test the hypotheses

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

This can be done via *Autocorrelation test*

Autocorrelation test

The autocovariance of the residuals is estimated as:

$$\hat{R}_e(\tau) = \frac{1}{N}\sum_{t=\tau}^{N} e(t)e(t-\tau)$$

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

$$N\frac{\sum_{k=1}^{m}\hat{R}_e^2(k)}{\hat{R}_e^2(0)} \to \chi^2(m)$$

Furthermore, the normalized autocovariance estimate is asymptotically Gaussian distributed

$$\sqrt{N}\frac{\hat{R}_e(\tau)}{\hat{R}_e(0)} \to \mathcal{N}(0,1)$$

A typical way of using the first test statistics for validation is as follows

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

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

$$N\frac{\sum_{k=1}^m \hat{R}_e^2(k)}{\hat{R}_e^2(0)} < \chi^2_\alpha(m) \quad \text{accept} \quad H_0$$

(m is often chosen from 5 up to $N/4$)

Cross Correlation test

The input and the residuals should be uncorrelated (no unmodeled dynamics)

$$R_{eu}(\tau) = \mathbf{E}\, e(t)u(t-\tau) = 0$$

- If the model is not an accurate representation of the system, one can expect $R_{eu}(\tau)$ for $\tau \geq 0$ is far from zero
- Indication of possible feedback in the input
- If $R_{eu}(\tau) \neq 0$ for $\tau < 0$ then there is output feedback in the input
- Use the normalized test quantity

$$x_\tau = \frac{\hat{R}_{eu}(\tau)^2}{\hat{R}_e(\tau)\hat{R}_u(0)}$$

for checking wheter the input and the residual are uncorrelated

For this purpose, introduce

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

$$r = \frac{1}{N}\sum_{t=1}^{N}\begin{bmatrix}u(t-\tau-1)\\ \vdots \\ u(t-\tau-m)\end{bmatrix}e(t)$$

where τ is a given integer and assume that $u(t) = 0$ for $t \leq 0$

Then we have

$$Nr^*[\hat{R}_e(0)\hat{R}_u]^{-1}r \longrightarrow \chi^2(m)$$

which can be used to design a hypothesis test

Numerical examples

The system that we will identify is given by

$$(1-1.5q^{-1}+0.7q^{-2})y(t) = (1.0q^{-1}+0.5q^{-2})u(t)+(1-1.0q^{-1}+0.2q^{-2})\nu(t)$$

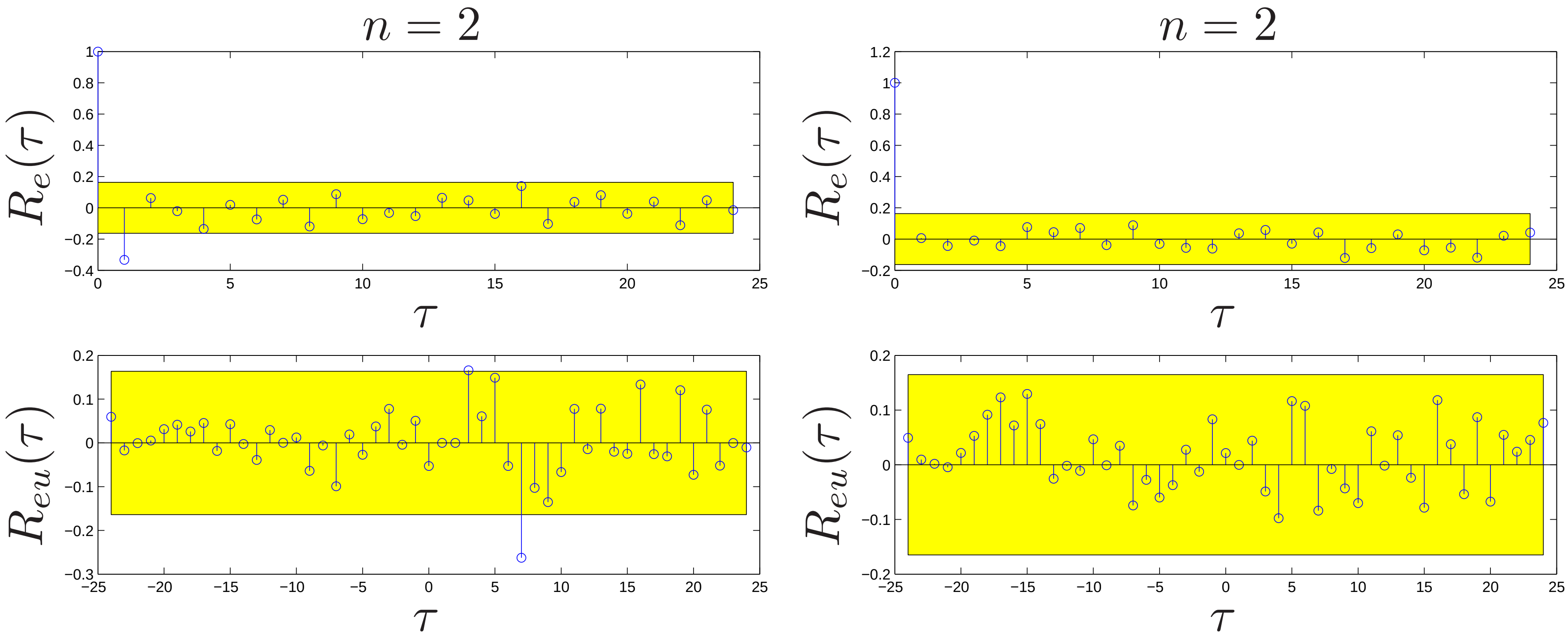
- $u(t)$ is binary white noise, independent of the white noise $\nu(t)$
- Generate two sets of data, one for estimation and one for validation
- Each data set contains data points of $N = 250$

Estimation:

- fitting ARX model of order n using the LS method
- fitting ARMAX model of order n using PEM

and vary n from 1 to 6

Example of residual analysis

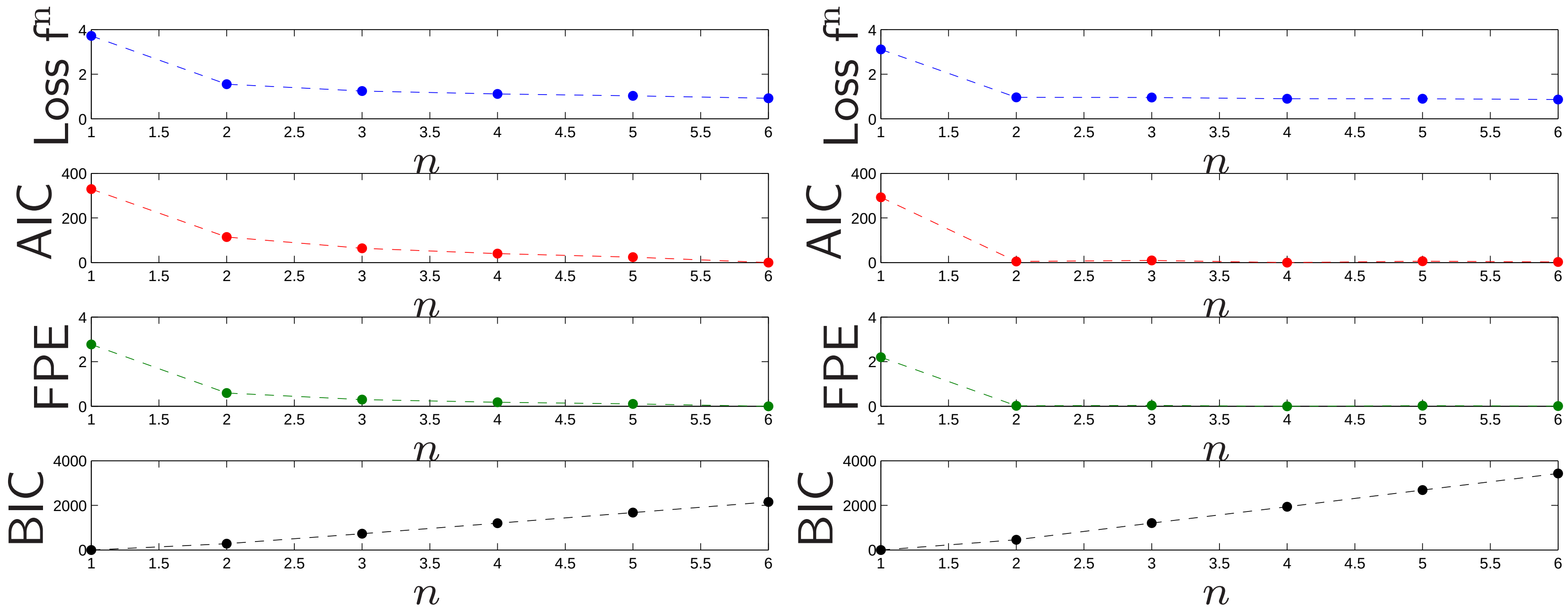


(*Left.* LS method *Right.* PEM)

- The significant correlation of e shows that e cannot be seen as white noise, or the noise model H is not adequate

- The significant correlation between e and u shows the dynamics G is not adequate

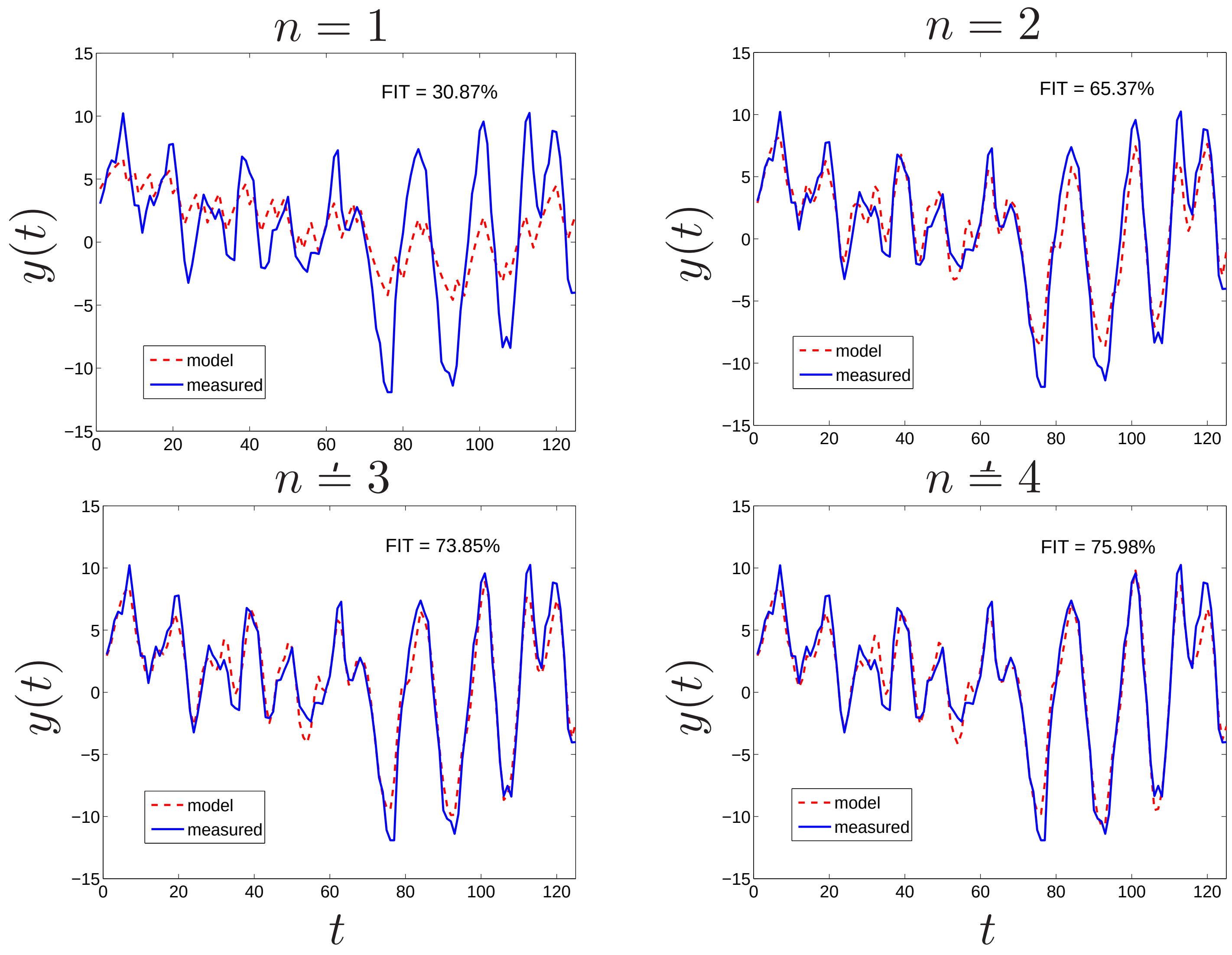
Example of model selection scores



(*Left.* LS method *Right.* PEM)

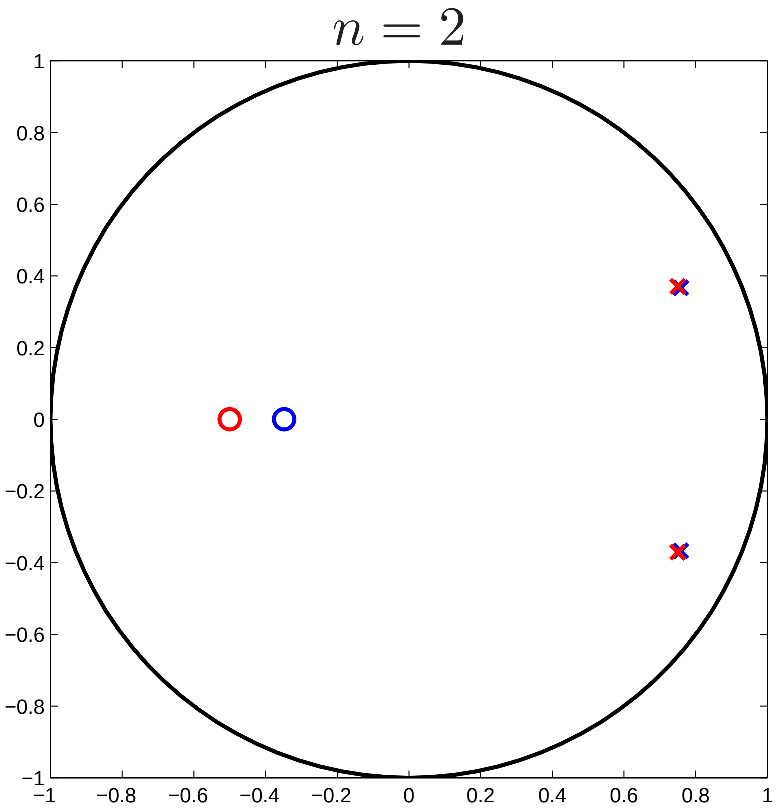
- AIC and FPE mostly pick higher models $(n = 4, 6)$
- BIC picks the simplest model
- All these scores decrease significantly at $n = 2$

Example of output prediction

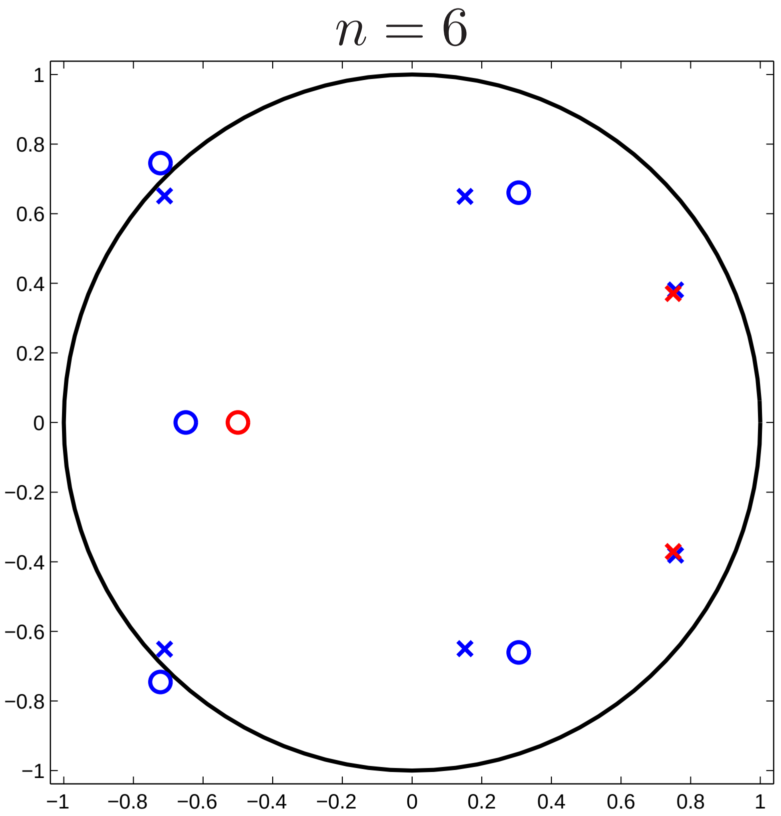


(estimated by the LS method and validated on a new data set)

Example of zero-pole location



$n = 2$



$n = 6$

- estimated by PEM, ○: zeros, ×: poles
- red: true system, blue: estimated models
- chance of zero-pole cancellation at higher order

Example of MATLAB commands

Suppose `theta` is an `idobject` obtained by using System Iden toolbox

- `armax`: estimate ARMAX models using PEM
- `iv4`: estimate ARX models using IVM
- `arx`: estimate ARX models using the LS method
- `resid`: residual analysis
- `compare`: compare the prediction with the measurement
- `zpplot`: Plots of zeros and poles
- `theta.EstimationInfo.LossFcn`: The value of loss function
- `theta.EstimationInfo.FPE`: The value of FPE

References

Chapter 11 in
T. Söderström and P. Stoica, *System Identification*, Prentice Hall, 1989

Chapter 16 in
L. Ljung, *System Identification: Theory for the User*, 2nd edition, Prentice Hall, 1999

Lecture on
Model Structure Determination and Model Validation, System Identification (1TT875), Uppsala University,
`http://www.it.uu.se/edu/course/homepage/systemid/vt05`

Chapter 7 in T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second edition, Springer, 2009