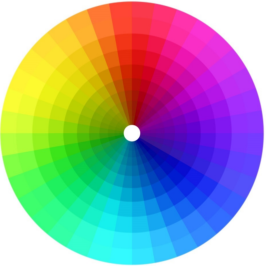
UPPSALA
UNIVERSITET

Order Selection



Per Mattsson

Systems and Control
Department of Information Technology
Uppsala University

2019-09-23

Summary from last lecture

- Alternatives to NLS, based on ARMA and covariance model
- HOYW, MUSIC, Min-Norm, ESPRIT
- Subspace methods using EVD/SVD
- Complicated derivation but easy to use (and implement)
- Need to select user parameters and model order

Today

- How to choose the model order for parametric methods?
- Heuristic approaches
- Information criteria
- Intuitive ARMA order selection

Parametric signal models

$$\boxed{y(t) = \sum_{k=1}^{n} \alpha_k e^{i(\omega_k t + \varphi_k)} + e(t)}$$

$$\boxed{A(z)y(t) = B(z)e(t)}$$

$$A(z) = 1 + a_1 z^{-1} + \ldots + a_n z^{-n}$$

$$B(z) = 1 + b_1 z^{-1} + \ldots + b_m z^{-m}$$

- We can use these signal models and estimate spectra by estimating the parameters (real valued)

Parametric signal models

$$\boxed{y(t) = \sum_{k=1}^{n} \alpha_k e^{i(\omega_k t + \varphi_k)} + e(t)}$$

$$\boxed{A(z)y(t) = B(z)e(t)}$$

$$A(z) = 1 + a_1 z^{-1} + \ldots + a_n z^{-n}$$
$$B(z) = 1 + b_1 z^{-1} + \ldots + b_m z^{-m}$$

- We can use these signal models and estimate spectra by estimating the parameters (real valued)

Remaining problem

What is n and $m \in \mathbb{N}$?

- How to estimate the **discrete** parameters?

Definitions

Refer to n as the model order, or rather, the number of parameters, and N as the number of *real-valued* samples

$$\theta \in \mathbb{R}^n, \qquad y \in \mathbb{R}^N$$

Definitions

Refer to n as the model order, or rather, the number of parameters, and N as the number of *real-valued* samples

$$\theta \in \mathbb{R}^n, \qquad y \in \mathbb{R}^N$$

For $\{y(t)\}_{t=1}^{N_s}$ complex-valued samples from the line spectra model

$$y(t) = \sum_{k=1}^{n_c} \alpha_k e^{i(\omega_k t + \varphi_k)} + e(t)$$

we have

$$N = 2N_s$$
$$n = 3n_c + 1$$

that is, both real and imaginary part of the data, and three parameters per component plus the noise variance are unknown

Rule of thumb

General

It is always *possible* to get a better model fit if we increase the model order ("increase flexibility").

- Infinite order is not better (or even possible)!
- Does not explain the underlying structure
- Fits to the random noise, giving random estimates (overfitting)

UPPSALA UNIVERSITET

Rule of thumb

General

It is always *possible* to get a better model fit if we increase the model order (“increase flexibility”).

- Infinite order is not better (or even possible)!
- Does not explain the underlying structure
- Fits to the random noise, giving random estimates (overfitting)

Heuristic approach (Occam’s razor)

We need to choose n high enough that the model gives a sufficient description of the data, while still keeping $n << N$ to get reliable (low variance) estimates.

In practice

Principle of parsimony idea

Increase the order as long as the error reduces **significantly**

- Subjective but reasonable
- An even lower model order might still be enough for your purpose
- What is significant?

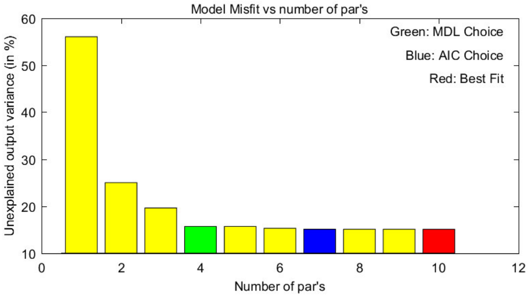
UPPSALA UNIVERSITET

In practice

Principle of parsimony idea

Increase the order as long as the error reduces **significantly**

- Subjective but reasonable
- An even lower model order might still be enough for your purpose
- What is significant?



Order selection rules

- Somehow *automatically* estimate n from y
- Many application specific methods (of limited applicability)
- **Here:** General rules associated with the Maximum Likelihood Method (MLM)

Order selection rules

- Somehow *automatically* estimate n from y
- Many application specific methods (of limited applicability)
- **Here:** General rules associated with the Maximum Likelihood Method (MLM)

Maximum likelihood

$p(y|\theta)$ is the probability of the data vector y given the model parameter vector θ.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ p(y|\theta) = \underset{\theta}{\operatorname{argmin}}\ -\ln\left(p(y|\theta)\right)$$

UPPSALA UNIVERSITET

Order selection rules

- Somehow *automatically* estimate n from y
- Many application specific methods (of limited applicability)
- **Here:** General rules associated with the Maximum Likelihood Method (MLM)

Maximum likelihood

$p(y|\theta)$ is the probability of the data vector y given the model parameter vector θ.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ p(y|\theta) = \underset{\theta}{\operatorname{argmin}}\ -\ln\left(p(y|\theta)\right)$$

- $p(y|\theta)$ is called the *likelihood function* (arbitrary noise distr.)

Order selection rules

- Somehow *automatically* estimate n from y
- Many application specific methods (of limited applicability)
- **Here:** General rules associated with the Maximum Likelihood Method (MLM)

Maximum likelihood

$p(y|\theta)$ is the probability of the data vector y given the model parameter vector θ.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ p(y|\theta) = \underset{\theta}{\operatorname{argmin}}\ -\ln\left(p(y|\theta)\right)$$

- $p(y|\theta)$ is called the *likelihood function* (arbitrary noise distr.)
- Intuitive – we want to maximize the probability that our model explains the data

Order selection rules

- Somehow *automatically* estimate n from y
- Many application specific methods (of limited applicability)
- **Here:** General rules associated with the Maximum Likelihood Method (MLM)

Maximum likelihood

$p(y|\theta)$ is the probability of the data vector y given the model parameter vector θ.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ p(y|\theta) = \underset{\theta}{\operatorname{argmin}}\ -\ln\left(p(y|\theta)\right)$$

- $p(y|\theta)$ is called the *likelihood function* (arbitrary noise distr.)
- Intuitive – we want to maximize the probability that our model explains the data
- Reduces to NLS for Gaussian data

Gaussian likelihood example

- Data-model, with parameters γ and noise variance σ^2:

$$y = f(\gamma) + e$$

Gaussian likelihood example

- Data-model, with parameters γ and noise variance σ^2:

$$y = f(\gamma) + e$$

- Likelihood (assuming Gaussian noise):

$$p(y|\theta) = \frac{1}{(2\pi)^{N/2}(\sigma^2)^{N/2}} e^{-\frac{\|y-f(\gamma)\|^2}{2\sigma^2}}$$

where $\theta = [\gamma,\ \sigma^2]^\top$ is the vector of all unknowns.

per.mattsson@it.uu.se

Gaussian likelihood example

- Data-model, with parameters γ and noise variance σ^2:

$$y = f(\gamma) + e$$

- Likelihood (assuming Gaussian noise):

$$p(y|\theta) = \frac{1}{(2\pi)^{N/2}(\sigma^2)^{N/2}} e^{-\frac{\|y-f(\gamma)\|^2}{2\sigma^2}}$$

where $\theta = [\gamma,\ \sigma^2]^\top$ is the vector of all unknowns.

- Log-likelihood:

$$-\ln(p(y|\theta)) =$$

Gaussian likelihood example

- Data-model, with parameters γ and noise variance σ^2:

$$y = f(\gamma) + e$$

- Likelihood (assuming Gaussian noise):

$$p(y|\theta) = \frac{1}{(2\pi)^{N/2}(\sigma^2)^{N/2}} e^{-\frac{\|y-f(\gamma)\|^2}{2\sigma^2}}$$

where $\theta = [\gamma,\ \sigma^2]^\top$ is the vector of all unknowns.

- Log-likelihood:

$$-\ln(p(y|\theta)) = \frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln(\sigma^2) + \frac{\|y - f(\gamma)\|^2}{2\sigma^2}$$

Gaussian likelihood example

- Data-model, with parameters γ and noise variance σ^2:

$$y = f(\gamma) + e$$

- Likelihood (assuming Gaussian noise):

$$p(y|\theta) = \frac{1}{(2\pi)^{N/2}(\sigma^2)^{N/2}} e^{-\frac{\|y-f(\gamma)\|^2}{2\sigma^2}}$$

where $\theta = [\gamma, \ \sigma^2]^\top$ is the vector of all unknowns.

- Log-likelihood:

$$-\ln(p(y|\theta)) = \frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln(\sigma^2) + \frac{\|y - f(\gamma)\|^2}{2\sigma^2}$$

- Minimizer:

$$\hat{\gamma} =$$

Gaussian likelihood example

- Data-model, with parameters γ and noise variance σ^2:

$$y = f(\gamma) + e$$

- Likelihood (assuming Gaussian noise):

$$p(y|\theta) = \frac{1}{(2\pi)^{N/2}(\sigma^2)^{N/2}} e^{-\frac{\|y-f(\gamma)\|^2}{2\sigma^2}}$$

where $\theta = [\gamma,\ \sigma^2]^\top$ is the vector of all unknowns.

- Log-likelihood:

$$-\ln(p(y|\theta)) = \frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln(\sigma^2) + \frac{\|y-f(\gamma)\|^2}{2\sigma^2}$$

- Minimizer:

$$\hat{\gamma} = \underset{\gamma}{\operatorname{argmin}}\ \|y - f(\gamma)\|^2$$

per.mattsson@it.uu.se

Gaussian likelihood example

- Data-model, with parameters γ and noise variance σ^2:

$$y = f(\gamma) + e$$

- Likelihood (assuming Gaussian noise):

$$p(y|\theta) = \frac{1}{(2\pi)^{N/2}(\sigma^2)^{N/2}} e^{-\frac{\|y-f(\gamma)\|^2}{2\sigma^2}}$$

where $\theta = [\gamma,\ \sigma^2]^\top$ is the vector of all unknowns.

- Log-likelihood:

$$-\ln(p(y|\theta)) = \frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln(\sigma^2) + \frac{\|y-f(\gamma)\|^2}{2\sigma^2}$$

- Minimizer:

$$\hat{\gamma} = \underset{\gamma}{\operatorname{argmin}}\ \|y-f(\gamma)\|^2 \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{N}\|y-f(\hat{\gamma})\|^2$$

Maximum a posteriori (MAP)

Hypothesis: H_n denotes that the model order is n
Bayes rule:

$$p(H_n|y) = \frac{p(y|H_n)p(H_n)}{p(y)}$$

where $p(H_n)$ is the *a priori* probability of H_n

Maximum a posteriori (MAP)

Hypothesis: H_n denotes that the model order is n
Bayes rule:

$$p(H_n|y) = \frac{p(y|H_n)p(H_n)}{p(y)}$$

where $p(H_n)$ is the *a priori* probability of H_n

MAP

Find the most probable order, given the data, through

$$\max_{n \in [1,\ \bar{n}]} p(y|H_n)p(H_n)$$

$p(y)$ is just a normalization factor independent of n

A few different rules

Derived from statistical reasoning and information theory (Maximum a posteriori or Kullback-Leibler information)

Four methods we will look at (listed by increasing “performance”)

- Akaike information criterion (AIC)
- Corrected Akaike information criterion ($\mathrm{AIC_c}$)
- Generalized information criterion (GIC)
- Bayesian information criterion (BIC)

See the book for derivations (beyond our scope).

A few different rules

Derived from statistical reasoning and information theory (Maximum a posteriori or Kullback-Leibler information)

Four methods we will look at (listed by increasing “performance”)

- Akaike information criterion (AIC)
- Corrected Akaike information criterion ($\mathrm{AIC_c}$)
- Generalized information criterion (GIC)
- Bayesian information criterion (BIC)

See the book for derivations (beyond our scope).

Several other criteria available too:

- Minimum description length (MDL)
- etc...

Statistical approach

Reasonable idea: Add a term to the fitting problem that depends on n, penalizing high order.

Family of selection rules

$$\underset{\theta_n, n}{\text{minimize}} \; -2\ln(p_n(y|\theta_n)) + \eta(n, N)n$$

where θ_n is used as a reminder that θ is of length n

The penalty coefficients $\eta(n, N)$ are given by

$$\begin{aligned} \text{AIC} &: \eta(n, N) = 2 \\ \text{AIC}_\text{c} &: \eta(n, N) = 2\frac{N}{N-n-1} \\ \text{GIC} &: \eta(n, N) = \nu \in [2,\ 6] \\ \text{BIC} &: \eta(n, N) = \ln N \end{aligned}$$

Penalty comparison

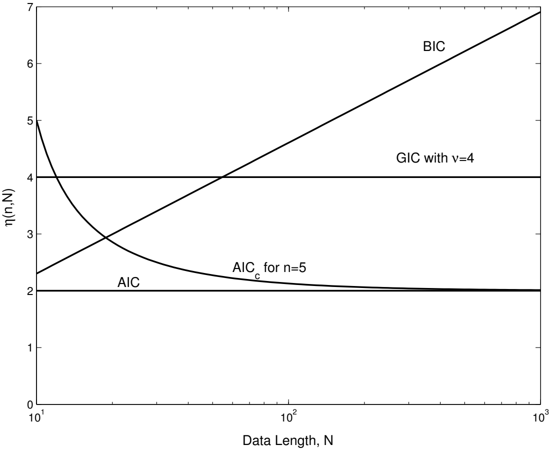


Figure C.1. Penalty coefficients of AIC, GIC with $\nu = 4$ $(\rho = 3)$, $\mathrm{AIC_c}$ (for $n = 5$), and BIC, as functions of data length N.

Practical considerations

- Hard to solve

$$\min_{\theta_n, n} -2\ln(p_n(y|\theta_n)) + \eta(n,N)n$$

- Assuming Gaussian noise, inserting the solution for fixed n

$$-2\ln(p_n(y|\hat{\theta}_n)) = N\ln(2\pi) + N + N\ln(\hat{\sigma}_n^2),$$

  where $\hat{\sigma}_n^2 = \frac{1}{N}\|y - f(\hat{\gamma}_n)\|^2$.
- Solution: Compute $\hat{\sigma}_n^2$ for many n, and choose the solution that minimize

$$N\ln(\hat{\sigma}_n^2) + \eta(n,N)n.$$

UPPSALA UNIVERSITET

Example: Line spectra

For some fixed order n_c of the complex-valued signal model, and the estimated parameters for that order, we have

$$\hat{\sigma}_{n_c}^2 = \frac{1}{N_s} \sum_{t=1}^{N_s} \left| y(t) - \sum_{k=1}^{n_c} \hat{\alpha}_k e^{i(\hat{\omega}_k + \hat{\varphi}_k)} \right|$$

which can be computed for every order n_c

per.mattsson@it.uu.se

Example: Line spectra

For some fixed order n_c of the complex-valued signal model, and the estimated parameters for that order, we have

$$\hat{\sigma}^2_{n_c} = \frac{1}{N_s} \sum_{t=1}^{N_s} \left| y(t) - \sum_{k=1}^{n_c} \hat{\alpha}_k e^{i(\hat{\omega}_k + \hat{\varphi}_k)} \right|$$

which can be computed for every order n_c

We can then compute, e.g. the AIC, as

$$\mathrm{AIC}(n_c) = 2N_s \ln(\hat{\sigma}^2_{n_c}) + 2(3n_c + 1)$$

and choose the order that minimizes the AIC

- We need to compute the error (MSE) for many model orders
- Then we can choose based on some information criteria

UPPSALA UNIVERSITET

Considerations

- Automatic order selection is possible
- Now we have several criteria, how do we choose?
  - Pick your favorite
  - Look at all of them to make a final decision
  - Combine the information based approaches
- Computational burden can be a problem
- Methods can “fail”
- An informed guess can still be better
- Non-parametric approaches avoids this problem (almost)

ARMA order selection

Heuristic for ARMA (or linear systems)

Reduce order if you have pole/zero cancellation, i.e., if there are estimated poles and zeros that overlap (more or less) they may not influence the result.

$$y(t) = \frac{B(z)}{A(z)}e(t) = \frac{\tilde{B}(z)(1 - kz^{-1})}{\tilde{A}(z)(1 - kz^{-1})}e(t) = \frac{\tilde{B}(z)}{\tilde{A}(z)}e(t)$$

where $\tilde{B}$ and $\tilde{A}$ have lower order

- Model specific approach (but quite general)
- Easy to use
- Intuitive

Useful functions

Custom functions implemented:

- `armaorder(mvec,sig2,N,nu)`
- `sinorder(mvec,sig2,N,nu)`

Usage:

- `mvec`: vector of number of sinusoids (or complex exponentials for complex valued data)
- `sig2`: vector mean square errors (that is, estimate of σ^2) for model orders given in mvec
- `N`: number of real-valued data points
- `nu`: GIC parameter (usually $\nu \in [2, 6]$, default=4)
- output: the model orders that minimizes the AIC, AICc, GIC, and BIC criterions

per.mattsson@it.uu.se

Summary

- Out of a selection of models that are sufficient for the application, choose the simplest one
- In general: try to choose $n << N$
- Look at the increase in performance (decrease in error) as a function of n
- BIC, GIC, AIC_c, AIC can give automatic guidance
- Study your model and simplify (e.g. pole/zero cancellation)

per.mattsson@it.uu.se

Summary

- Out of a selection of models that are sufficient for the application, choose the simplest one
- In general: try to choose $n << N$
- Look at the increase in performance (decrease in error) as a function of n
- BIC, GIC, AIC_c, AIC can give automatic guidance
- Study your model and simplify (e.g. pole/zero cancellation)

In the end, try several things to make yourself comfortable with a certain choice of n