System Identification

6.435

SET 12

- Identification in Practice
- Error Filtering
- Order Estimation
- Model Structure Validation
- Examples

Munther A. Dahleh

"Practical" Identification

- Given: $Z^N = \{y(t), u(t); t \leq N\}$

- <u>Want</u>

    1) a model for the plant

    2) a model for the noise

    3) an estimate of the accuracy

- choice of the model structure

    flexibility   parsimony

- What do we know?

  We know methods for identifying “models” inside a “priori” given model structures.

- How can we use this knowledge to provide a model for the plant, the process noise, with reasonable accuracy.

Considerations

- Pre-treatment of data
  - Remove the biase (may not be due to inputs)
  - Filter the high frequency noise
  - Outliers ……
- Introduce filtered errors. Emphasize certain frequency range. (The filter depends on the criterion).

- Pick a model structure (or model structures)
  - Which one is better?
  - How can you decide which one reflects the real system?
  - Is there any advantage from picking a model with a large number of parameters, if the input is “exciting” only a smaller number of frequency points?

- What are the important quantities that can be computed directly from the data (inputs & outputs), that are important to identification?

Pre-treatment of Data

- Removing the biase

$$\boxed{Ay(t) = Bu(t) + v(t)}$$

  - If $Ev(t) = 0$, then the relation between the static input $u(t) = \bar{u}$ and output $y(t) = \bar{y}$ is given by

    $$A(1)\bar{y} = B(1)\bar{u}$$

  - The static component of $y(t) = \bar{y}$ <u>may not</u> be entirely due to $\bar{u}$, i.e. the noise might be biased.

– Method I: Subtract the means:

Define $$\bar{y} = \frac{1}{N}\sum_{t=1}^{N} y^m(t) \qquad [y^m = \text{meas. data}]$$

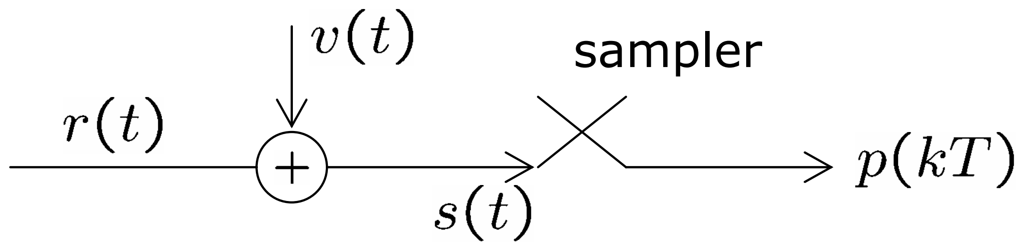
$$\bar{u} = \frac{1}{N}\sum_{t=1}^{N} u^m(t)$$

New Data: $$y(t) = y^m(t) - \bar{y}$$
$$u(t) = u^m(t) - \bar{u}$$

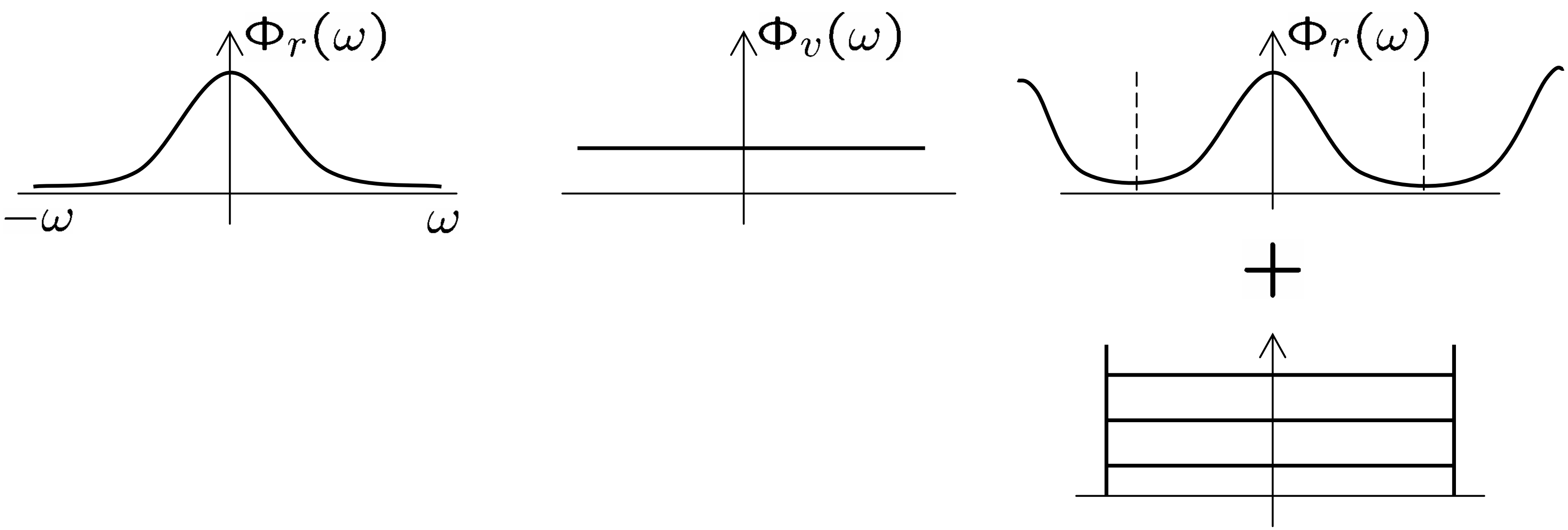
– Method II: Model the offset by an unknown constant α, and estimate it.

$$m: \quad Ay(t) = Bu(t) + \alpha + v(t)$$

- High Frequency disturbances in the data record.
  - "High" means above the frequency of interest.
  - Related to the choice of the sampling period.



$$T < \frac{1}{2W}$$



- Without an anti-aliasing filter, high frequency noise is folded to low frequency.

- $s(t) \longrightarrow \boxed{\boldsymbol{W}_{\text{anti-alias}}} \xrightarrow{\tilde{s}(t)} \boxed{\text{Sample}} \xrightarrow{\tilde{s}(kT)}$

- high frequency noise depends on:
  a) high frequency noise due to $v(t)$
  b) aliasing.

- Problem occurs at both the inputs and outputs.

$$y_F = Ly \qquad\qquad \boldsymbol{L} \text{ – LTI LP filter.}$$
$$u_F = Lu$$

- $m: \quad A(t)y_F = B(t)u_F + v(t) \qquad [v = He]$

equivalently $$A(t)y = B(t)u + \frac{1}{L}v(t)$$

i.e. multiply the noise filter by $\frac{1}{L}$

- Outliers, Bursts
  - Either erroneous or high-disturbed data point.
  - Could have a very bad effect on the estimate.
  - <u>Solution</u>:

    a) Good choice of a criterion (Robust to changes)

    b) Observe the residual spectrum. Sometimes it is possible to determine bad data.

    c) Remove by hand!!! Messing up with real data.

    d) Failure-detection using hypothesis testing or statistical methods. (Need to define a threshold).

Role of Filters: Affecting the Biase Distribution

- $m_1 : \quad Gu + H_1 e$

  $y_F = Ly$

  $u_F = Lu$

  $$\Rightarrow \quad m : \quad y(t) = Gu + He \qquad H = \frac{1}{L} H_1$$

- Frequency domain interpretation of parameter estimation:

$$\theta^* = \text{argmin} \int_{-\pi}^{\pi} \frac{\Phi_{ER}(\omega, \theta)}{\left| H\left(e^{i\omega}, \theta\right)\right|^2} d\omega$$

$$\Phi_{ER} = \left| G_o - G\left(e^{i\omega}, \theta\right)\right|^2 \Phi_u(\omega) + \Phi_v(\omega)$$

If $\theta = \begin{bmatrix} \rho \\ \eta \end{bmatrix}$: independently parametrized model structure

$$\rho^* = \underset{\rho}{\operatorname{argmin}} \int_{-\pi}^{\pi} \left| G_o\left(e^{i\omega}\right) - G\left(e^{i\omega}, \theta\right) \right|^2 Q(\omega, \eta^*) d\omega$$

$$Q(\omega, \eta^*) = \frac{\Phi_u}{\left| H\left(e^{i\omega}, \eta^*\right) \right|^2}$$

$$\eta^* = \operatorname{argmin} \int_{-\pi}^{\pi} \left| \frac{1}{N\left(e^{i\omega}, \rho^*\right)} - \frac{1}{H\left(e^{i\omega}, \eta\right)} \right|^2 d\omega$$

$$Q_{ER}(\omega, \rho^*) = \lambda^* N N^*$$

- Heuristically, θ is chosen as a compromise between minimizing the integral of $\left|G_o - G\left(e^{i\omega}, \theta\right)\right|^2 Q(\omega, \theta^*)$ and matching the error spectrum Φ_{ER}.

- $Q(\omega, \theta^*) = \dfrac{\Phi_u}{\left|H\left(e^{i\omega}, \theta^*\right)\right|^2}$ weighting function
  - Input spectrum
  - Noise spectrum

- With a pre-filter: $Q(\omega, \theta^*) \rightarrow \left|L\left(e^{i\omega}\right)\right|^2 Q(\omega, \theta)$

- Can view the pre-filters as weighting functions to emphasize certain frequency ranges. This interpretation <u>may not</u> coincide with "getting rid of high frequency components of the data".

- Depending on the criterion, the choice of L can be different.

OE Model Structures

- $m : Gu(t) + e(t) \qquad H = 1$

$$\rho^* = \text{argmin} \int_{-\pi}^{\pi} \left| G_o\left(e^{i\omega}\right) - G\left(e^{i\omega}, \theta\right) \right|^2 \Phi_u(\omega) d\omega$$

- If G_o rolls off, then as long as $G\left(e^{i\omega}, \theta\right)$ is small around $\omega \simeq \pi$ the contribution of the criterion $\left| G_o\left(e^{i\omega}\right) - G\left(e^{i\omega}, \theta\right) \right|$ will be very small.

- If $\Phi_u = 1$, we expect $G\left(e^{i\omega}, \theta\right)$ to match G_o much better at low-frequency.

- Example (book)

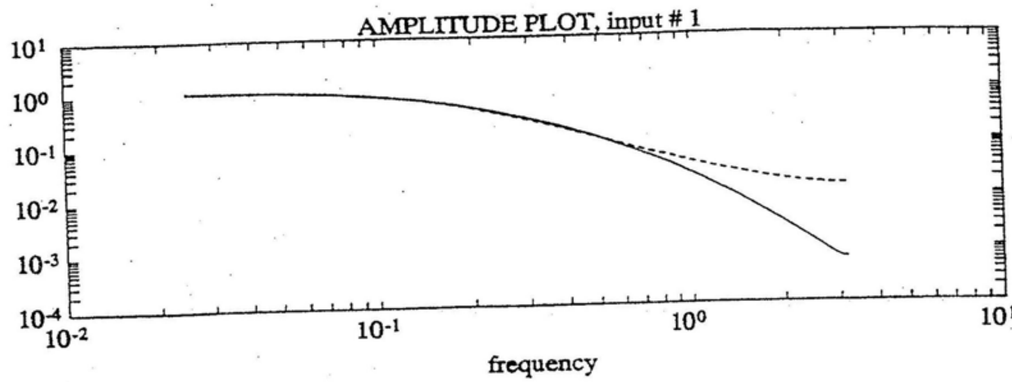
$$\delta : G_o(q) = \frac{0.001q^{-2}(10 + 7.4q^{-1} + 0.924q^{-2} + 0.1764q^{-3})}{1 - 2.14q^{-1} + 1.553q^{-2} - 0.4387q^{-3} + 0.042q^{-4}}$$

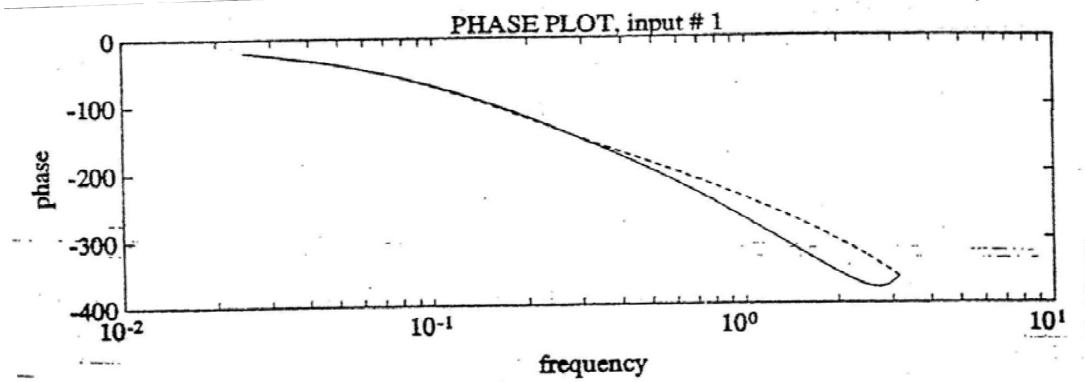
No noise.

$\Phi_u = 1$ PSRB

OE: $\hat{y} = G\left(e^{i\omega}, \theta\right) u$

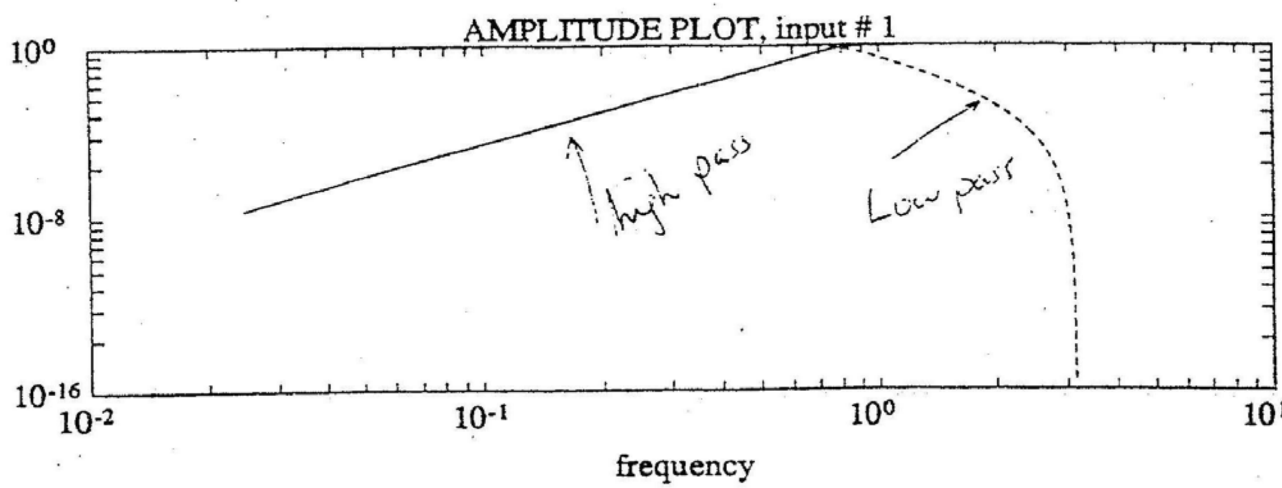
$$G = \frac{b_1 q^{-1} + b_2 q^{-2}}{1 + f_1 q^{-1} + f_2 q^{-2}}$$

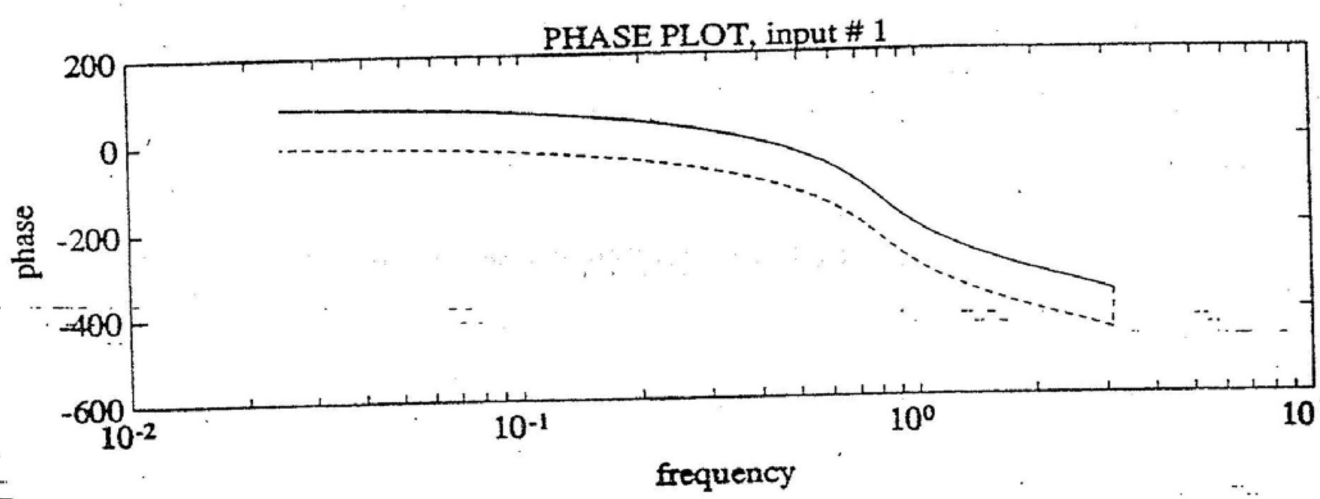


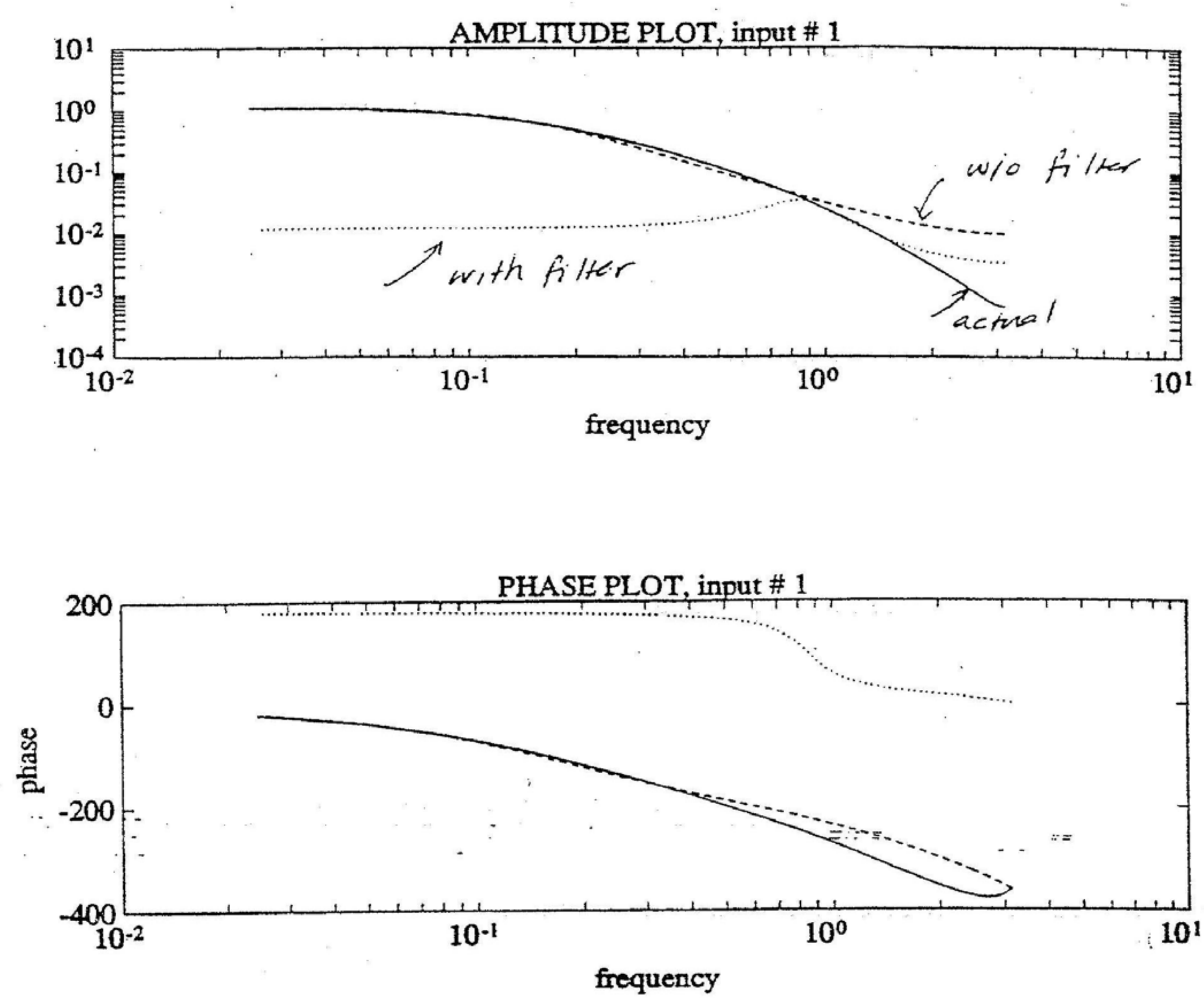


- Good match at low frequency.
  Not as good at high frequency.

- Introduce a high-pass filter (5^{th} order
  Butterworth filter, cut-off freq = 0.5 rad/sec.


AMPLITUDE PLOT, input # 1
10⁰
10⁻⁸
10⁻¹⁶
10⁻²
10⁻¹
10⁰
10¹
frequency
high pass
Low pass


PHASE PLOT, input # 1
200
0
-200
-400
-600
phase
10⁻²
10⁻¹
10⁰
10¹
frequency


AMPLITUDE PLOT, input # 1
10¹
10⁰
10⁻¹
10⁻²
10⁻³
10⁻⁴
10⁻²
10⁻¹
10⁰
10¹
frequency
w/o filter
with filter
actual
PHASE PLOT, input # 1
200
0
-200
-400
phase
10⁻²
10⁻¹
10⁰
10¹
frequency

ARX Model Structure

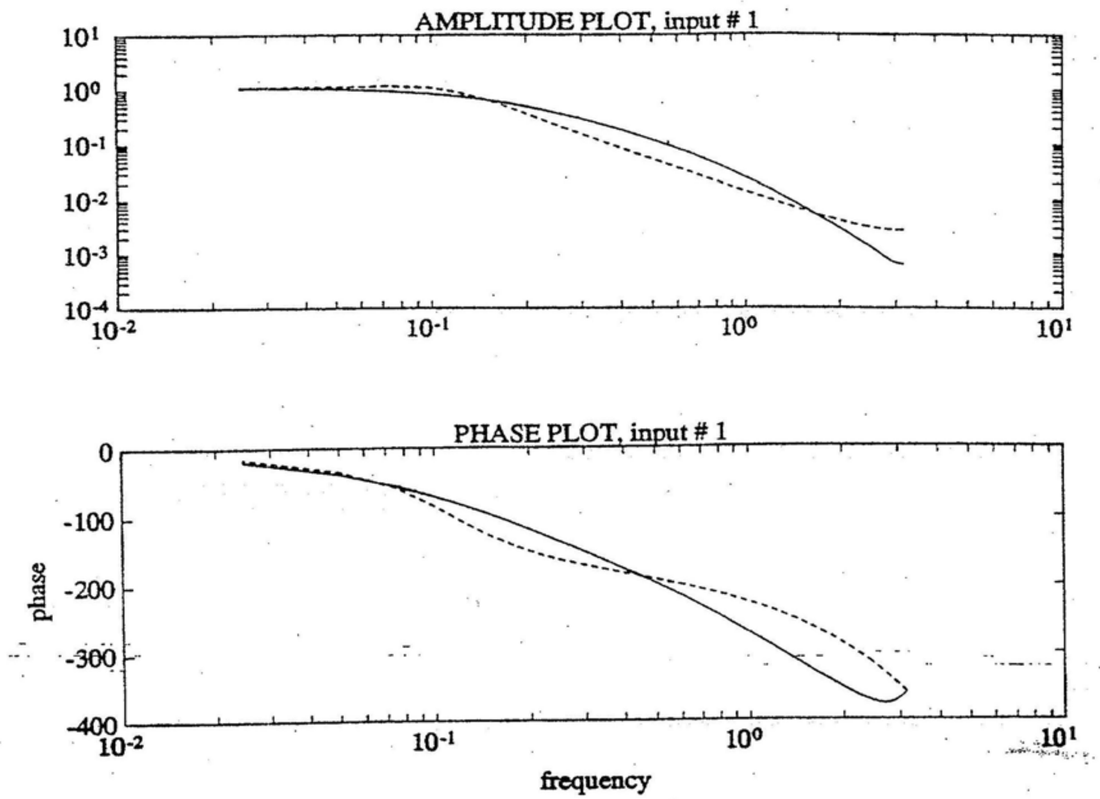
- $G = \dfrac{B}{A} \qquad H = \dfrac{1}{A}$ not independently-parametrized.

$$\theta^* = \text{argmin} \int_{-\pi}^{\pi} \left|G_o\left(e^{i\omega}\right) - G\left(e^{i\omega},\theta\right)\right|^2 \left|A\left(e^{i\omega},\theta\right)\right|^2 \Phi_u + \cancelto{=0}{\Phi_v} |A|^2$$

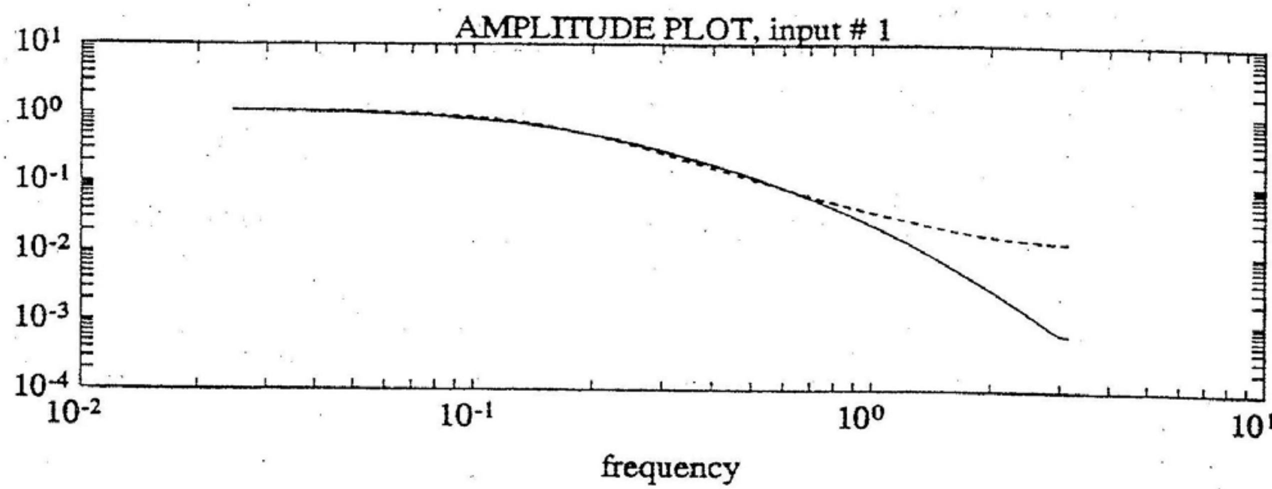
- If G_o rolls off, $B_o/A_o = G_o$ and A_o is large at high frequency. If $A\left(e^{i\omega},\theta\right)$ looks like A_o, then it will emphasize the high frequency part of the criterion.
- Conclusions are not as transparent in the noisy case. However, it is in general true for large (SNR).
- Same example

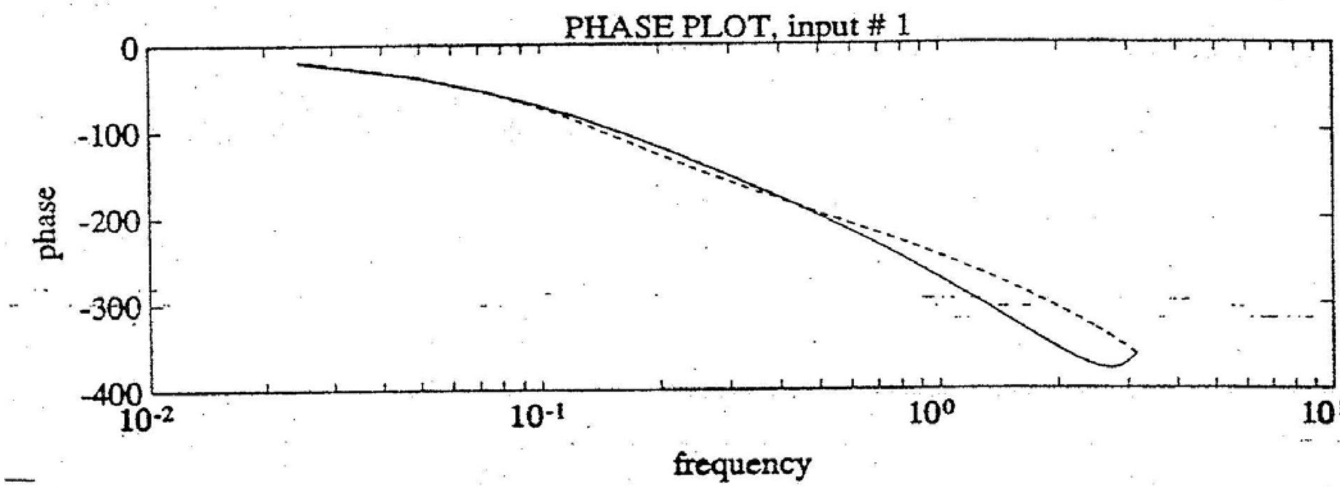
$$m: \quad Ay = Bu + e$$

and the frequency response of $G\left(e^{i\omega},\theta^*\right)$ is:



- Not a very good match at low frequency.
- Better than OE at high frequency.
- Can change this through a pre-filter. (5^{th} order Butterworth, lowpass with cut-off frequency = 0.5)


AMPLITUDE PLOT, input # 1
10¹
10⁰
10⁻¹
10⁻²
10⁻³
10⁻⁴
10⁻²
10⁻¹
10⁰
10¹
frequency


PHASE PLOT, input # 1
0
-100
-200
-300
-400
phase
10⁻²
10⁻¹
10⁰
10¹
frequency

– Better low frequency fit.

– Another interpretation

$$y = \frac{B}{A}u + \frac{1}{A}e \qquad \left|\frac{1}{A}\right|$$ small at high frequency.

low frequency

<u>Filters</u>:

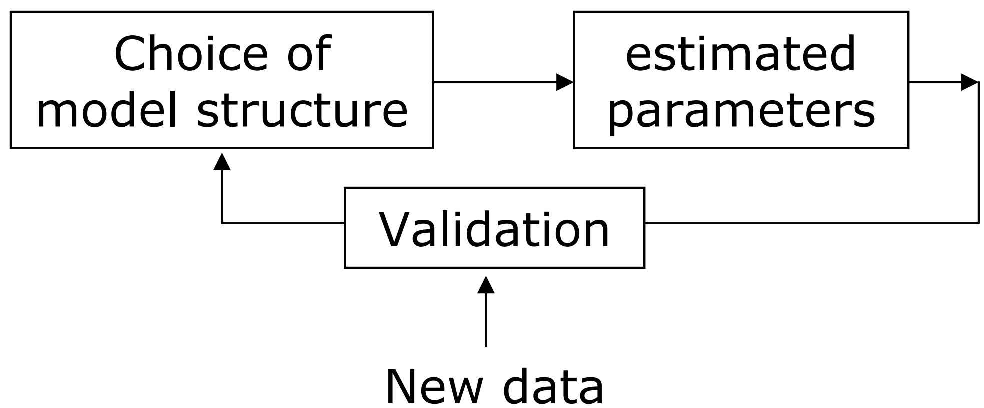
$$y = \frac{B}{A}u + \frac{L}{LA}e$$ high frequency if L – low pass.

Conclusions

- Pre filters can be viewed as “design” parameters as well as the “standard” interpretation for noise reduction.
- Pre-treatment of the “data” is quite valuable, however should be done with “care”.
- Sampling can be quite tricky. Need to estimate the bandwidth of the system.

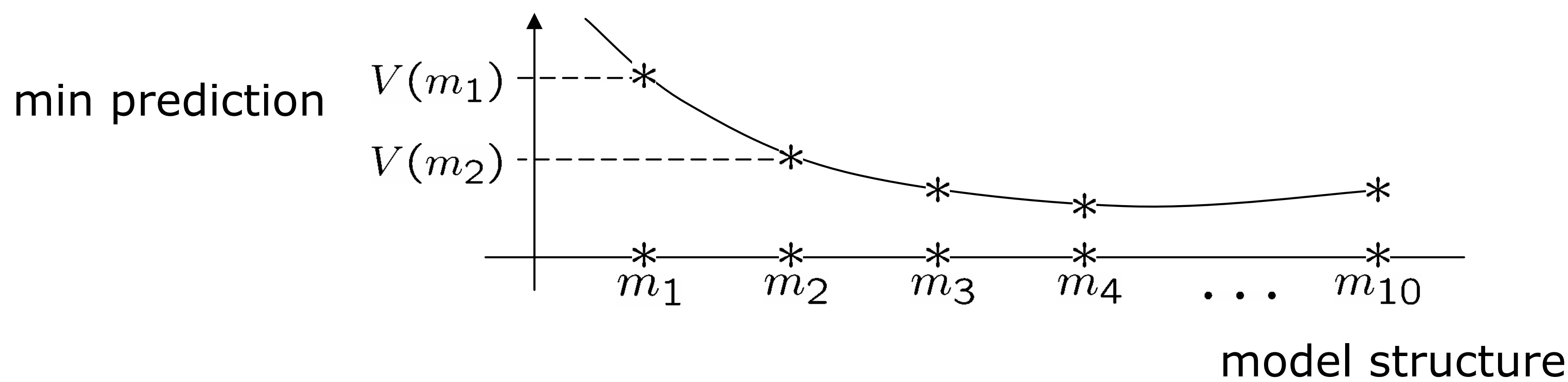
Model Structure Determination

- Flexible vs Parsimony
- Lots of trial and error. Usually more than one “experiment” is available.



- Theme: Fit the data with the least “complex” model structure. Avoid over-fitting which amounts to fitting noise.
- Better to compare similar model structures, although it is necessary to compare different structures at the end.

- Available data dictates the possible model structures and their dimensions.
- Can noise help identify the parameters?
- How “bad” is the effect of noise? Consistency in general is guaranteed for any SNR. What does that mean?
- Is there a rigorous way of comparing different model structures?



- Akaike’s Information Theoretic Criterion (AIC).
- Akaike’s final prediction error critierion (FPE).

Order Estimation

- Qualitative
  - Spectral estimate
  - Step response if available. Otherwise, step response of spectral estimate.
- Quantitative
  - Covariance matrices
  - Information matrix $E\left(\Psi(t,\theta)\Psi^T(t,\theta)\right)$
  - Residual-input correlation
  - Residual whiteness

$$\vdots$$

- All methods are limited by the input used.

Covariance Matrix

$$m: \overset{n}{A} y = \overset{n}{B} u + v \qquad \phi_s(t) = (-y(t-1), \ldots, -y(t-s), u(t-1), \ldots, u(t-s))$$

- Two basic results
  - $v = 0 \quad \bar{E}\phi_n(t)\phi_n^T(t) > 0 \Leftrightarrow$ $\boldsymbol{u}$ is p.e of order 2$\boldsymbol{n}$ and $\boldsymbol{A}$, $\boldsymbol{B}$ are coprime
  - $v \neq 0 \quad \bar{E}\phi_n(t)\phi_n^T(t) > 0 \Leftrightarrow$ $\boldsymbol{u}$ is p.e of order $\boldsymbol{n}$

    white or persistent
- To determine $\boldsymbol{n}$, obtain estimates of $\bar{E}\phi_s(t)\phi_s^T(t)$

Case 1: u is $\boldsymbol{WN}$, v is $\boldsymbol{WN}$

Increase s until $\bar{E}\phi_s\phi_s^T(t)$ is "singular".

$$\bar{E}\phi_s(t)\phi_s^T(t) \simeq \frac{1}{N}\sum_{t=1}^{N}\phi_s(t)\phi_s^T(t) = \widehat{R}_s$$

Use "SVD", robust rank tests.

$\det(\widehat{R}_s) \stackrel{?}{=} 0$, observe a sudden drop in the rank.

Case 2: u is p.e of order n_u, Noise is white.

If $s > n_u \Rightarrow \bar{E}\phi_s(t)\phi_s^T(t)$ is singular.

you really cannot estimate the order of the system if it is larger than n_u .

<u>Case 3</u>: $\boldsymbol{u}$ is p.e of order n_u, Noise free case.

$n > \frac{n_u}{2}$ cannot be determined. Not a likely hypothesized model structure.

$$\widehat{R}_s = \frac{1}{N}\sum_{t=1}^{N} \phi_s(t)\phi_s^T(t)$$

is a bad estimate of $\bar{E}\phi_s\phi_s^T(t)$. Of course $\boldsymbol{N}$ is fixed (data length).

"Enhanced criterion"

$$\widehat{\widehat{R}}_s = \widehat{R}_s - \underbrace{\widehat{\sigma} R_v}_{\text{estimated noise contribution.}}$$

- If noise level is high, use an instrumental variable

$$\xi_s(t) = [u(t-1), \ldots, u(t-2s)].$$

test: $\boxed{\text{rank} \quad \bar{E}\xi_s(t){\phi_s}^T(t)}$

- If u is p.e , then generically

$$\bar{E}\xi_s{\phi_s}^T > 0 \quad \Leftrightarrow \quad s \leq n \quad , \quad n \text{ order of } A, B$$

- Other tests:

Estimates of

$$\bar{E}u(t)\varepsilon(t,\theta) \quad \simeq 0$$

$$\bar{E}\varepsilon(t,\theta)\varepsilon(t-k,\theta) \quad k \neq 0 \quad \simeq 0$$

$$\bar{E}\Psi(t,\theta)\Psi^T(t,\theta) \quad \text{non singular}$$

Examples

- δ unknown. Study possible conclusions for different experiments and different SNR.
- Model structure

  $$m: \quad Ay = Bu + v$$

- Inputs
  - $\boldsymbol{WN}$
  - $\cos \omega_o \quad w_o = 2\pi/1000$
  - $\cos(2\pi/1000) \quad , \quad \cos 20 \cdot (2\pi/1000)$

Can determine from the spectrum or u (or simply FFT).

- SNR:

$$\lambda = 1$$
$$\lambda = 0.1$$
$$\lambda = 0.01$$

- All examples, you can access both the inputs and outputs.

First Experiment

$u = WN$

$\lambda = 1 \qquad SNR \simeq 1$

Test for model order:

$$\bar{E}\phi_s\phi_s^T(t) \cong \frac{1}{N}\sum_{t=1}^{N}\phi_s(t)\phi_s^T(t) = \widehat{R}_s$$

From data $\widehat{R}_s$ is singular for $s \geq 2$

$\det \widehat{R}_s \simeq 0$ (noticed a sudden drop)

$\rightarrow$ Estimated system

$$a_1 = -1.44 \qquad a_2 = 0.498$$

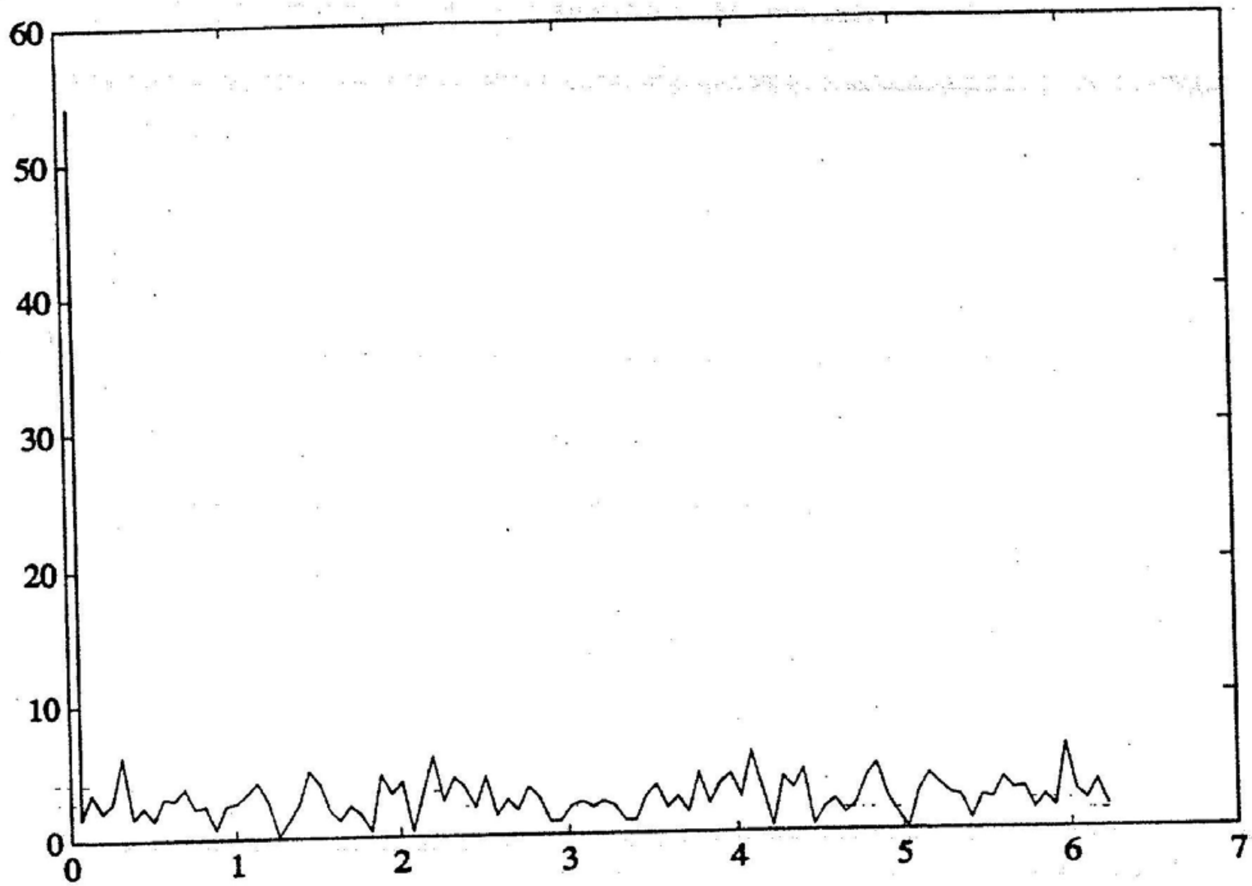
$$b_1 = 1.0221 \quad b_2 = 0.5239$$

$\rightarrow$ cov $\widehat{\theta}_N$ small

(ARX)

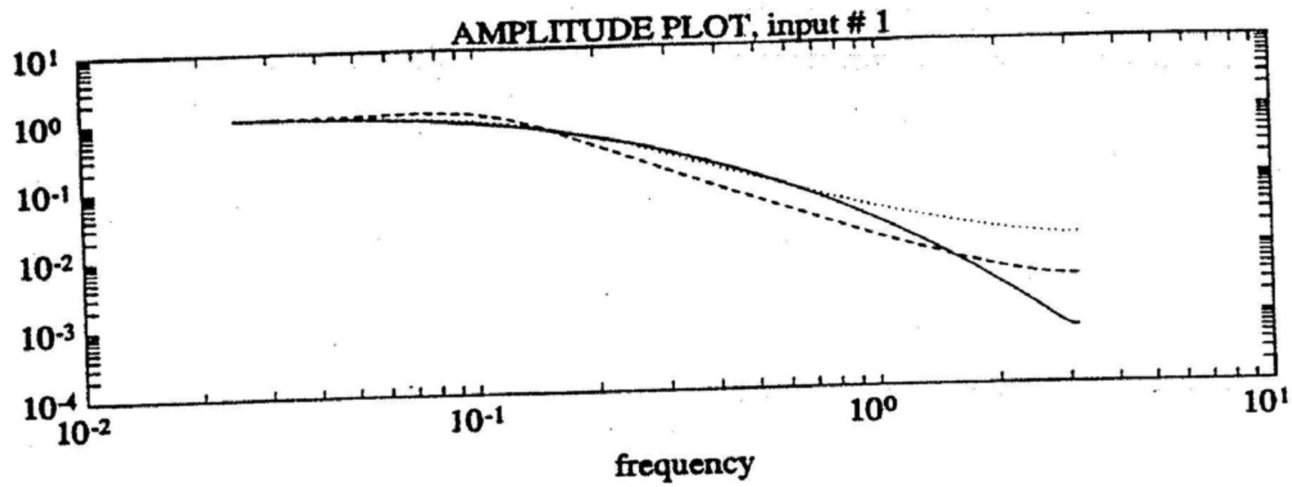
Comment: If m is the correct structure γ, the above are "good" estimates of the parameters.

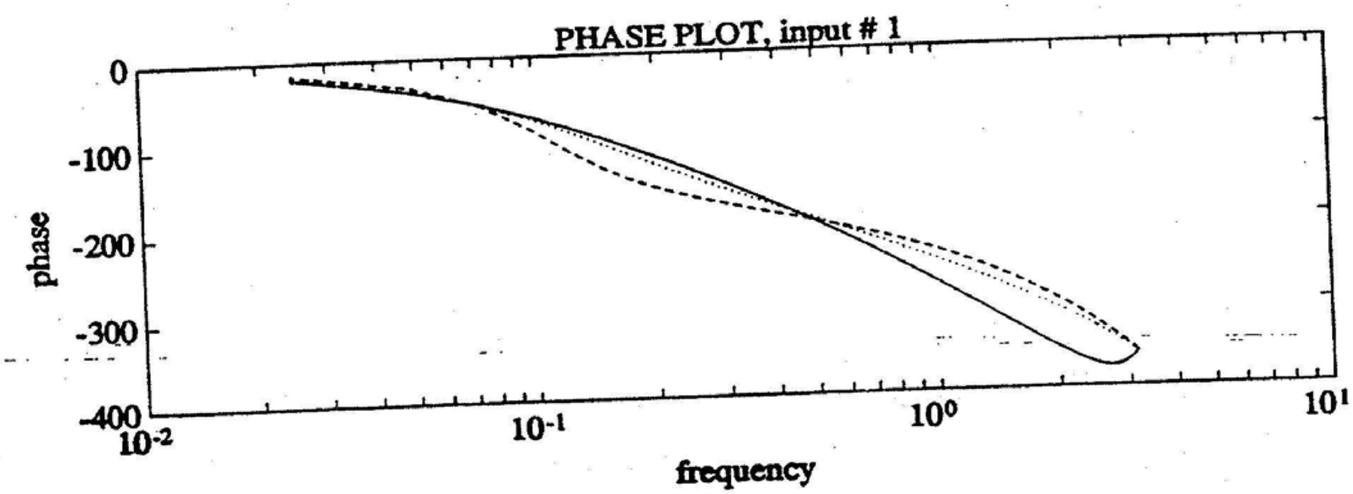
Plot shows different SNR.



FFT of the input

random


AMPLITUDE PLOT, input # 1
10^1
10^0
10^-1
10^-2
10^-3
10^-4
10^-2
10^-1
10^0
10^1
frequency


PHASE PLOT, input # 1
0
-100
-200
-300
-400
phase
10^-2
10^-1
10^0
10^1
frequency

ARX (ARX)

Second Experiment

$u = \cos w_o t$

u is p.e of order 2.

SNR	a_1	a_2	b_1	b_2	cov $\hat{\theta}_N$	det $\hat{R}$
1	-0.0014	0.0005	-0.0520	0.054	high	$2.4\text{x}10^{-6}$
0.1	-1.38	0.479	1.255	0.356	high	$3.03\text{x}10^{-7}$
0.01	-1.398	0.4885	0.9631	0.5437	low	$2.67\text{x}10^{-7}$

Theoretical Analysis:

Data is informative $\bar{E}\phi_2\phi_2^T > 0$ (although det is small)

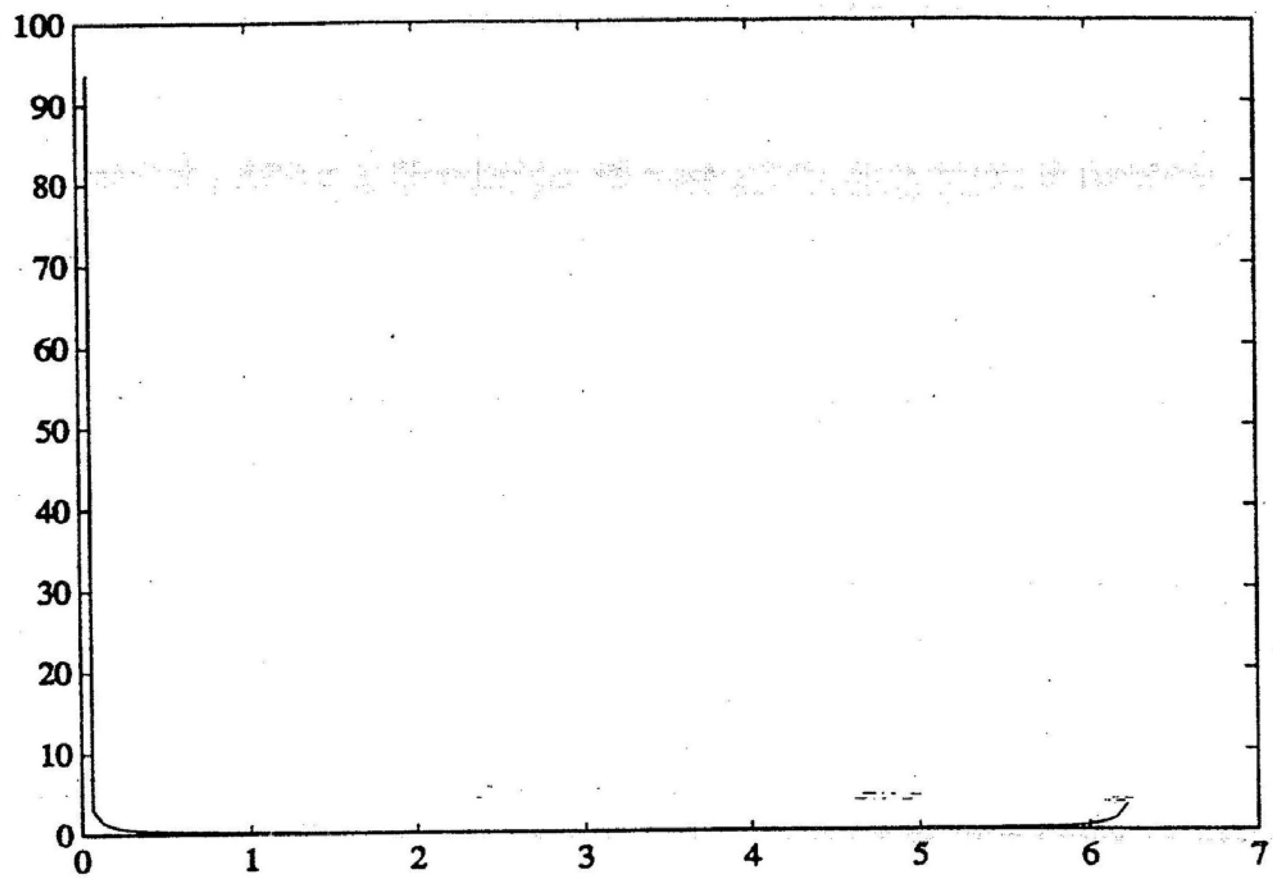
For $\cos w_o t$, as $N \to \infty$

$$\widehat{\theta}_N \to \widehat{\theta}_o^* \text{ regardless of } \lambda$$

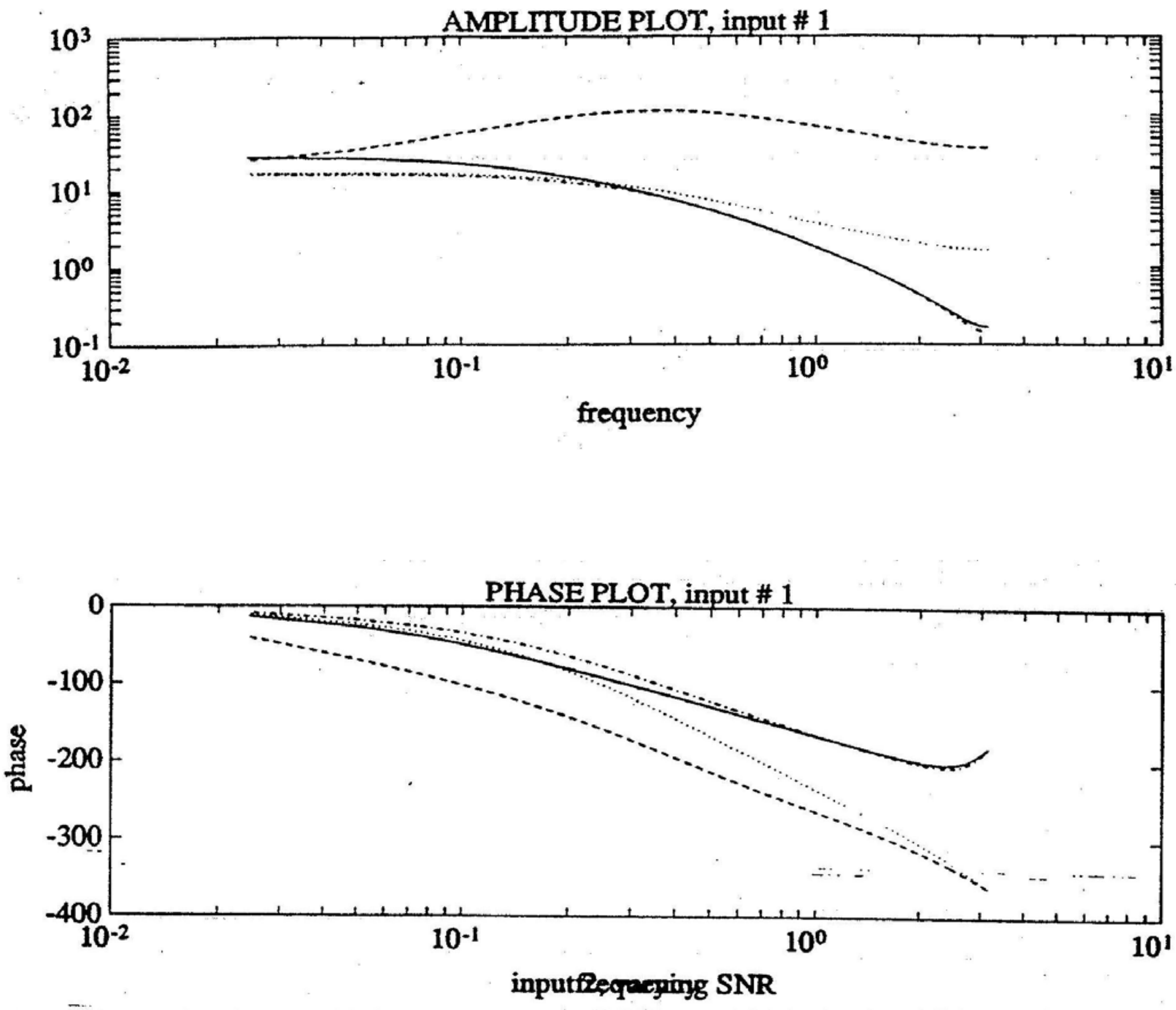
$N = 100$, the estimates were quite bad for $\lambda = 1$ in comparison to $\boldsymbol{WN}$ inputs.

Even though noise "helps" in obtaining asymptotic convergence (through providing excitation), it is not helpful for finite-data records, since its effect cannot be averaged. Accuracy depends on

$$\lambda\left(\bar{E}\Psi\Psi^T\right)^{-1}$$



FFT of second input



$u = \text{rand}$ λ = 1

$u = \cos w_o t$ λ = 1

λ = 0

λ = 0

ARX (with ARX plant)

2nd input

Third Experiment

$u = \cos w_o t + \cos 20 w_o t$

$\boldsymbol{u}$ is p.e of order 4.

SNR	a_1	a_2	b_1	b_2	cov $\widehat{\theta}_N$	det $\widehat{R}$
1	-1.475	0.5317	2.38	-1.134	high	0.25
0.1	-1.39	0.478	1.186	0.333	smaller	0.07
0.01	-1.395	0.4862	1.022	0.487	low	0.069

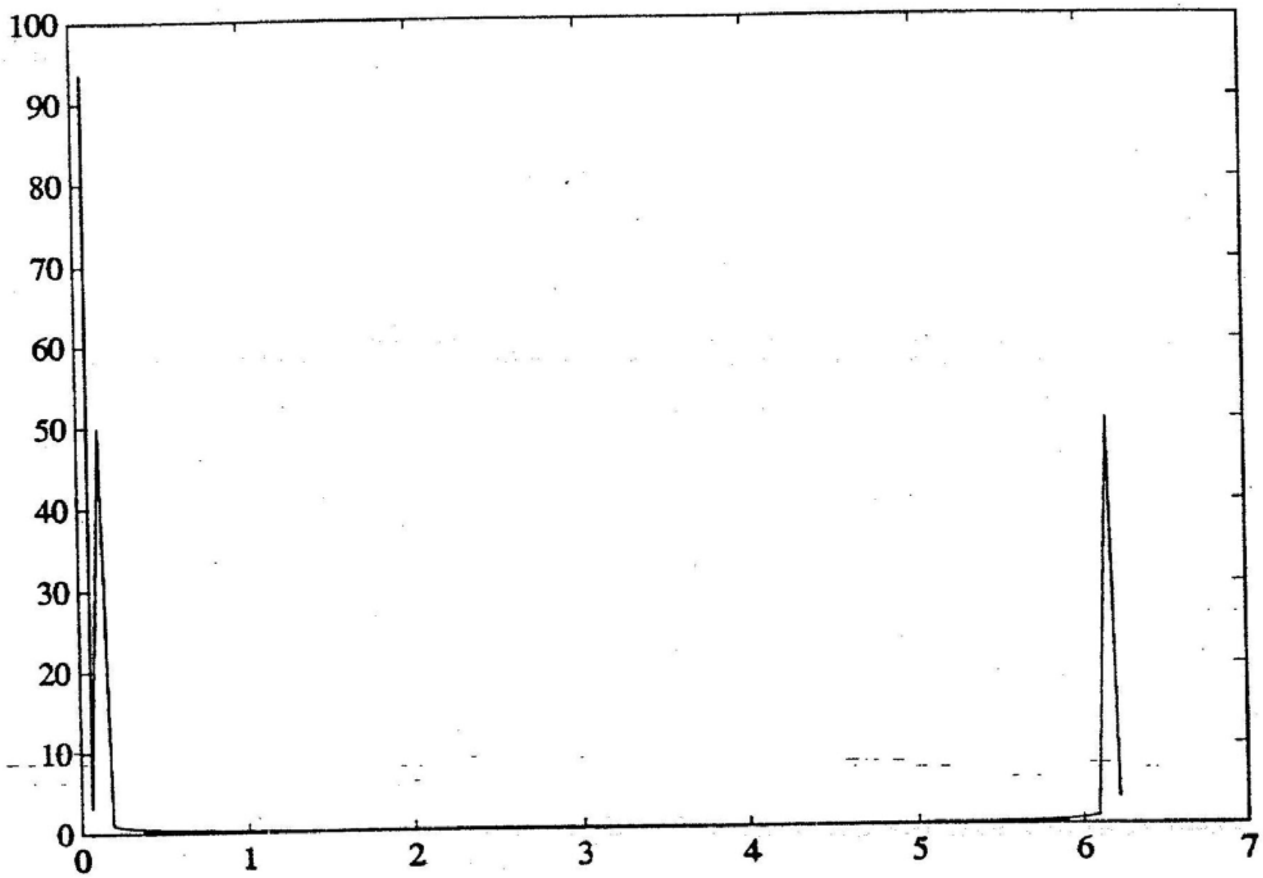
Theoretical Analysis:

n : use det $\hat{R}_s$ test $\Rightarrow$ $n \leq 2$

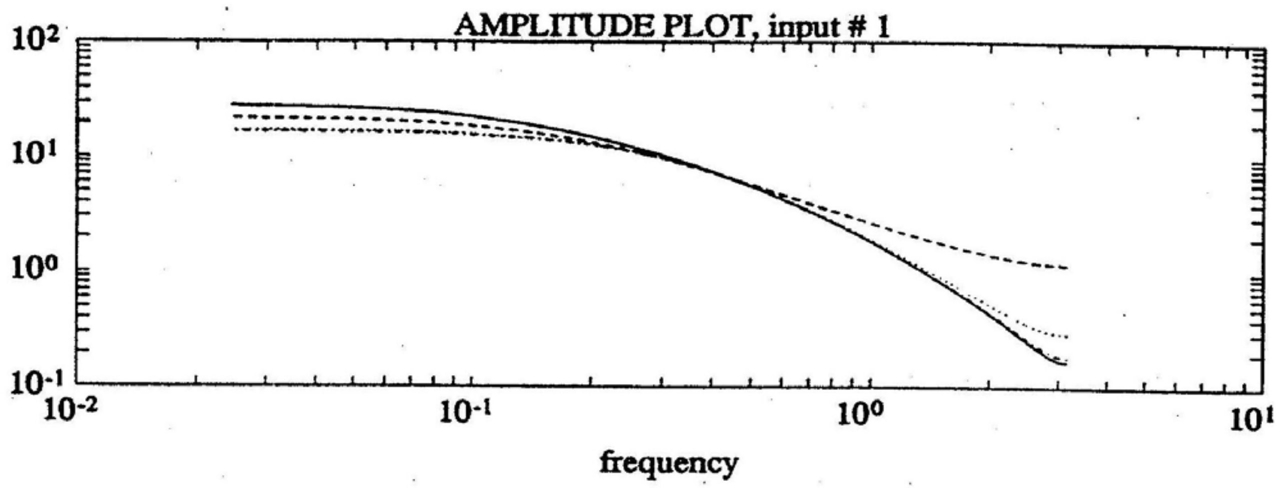
Data is informative w. r. to $\overset{\swarrow 2}{A} y = \overset{\swarrow 2}{B} u + e$

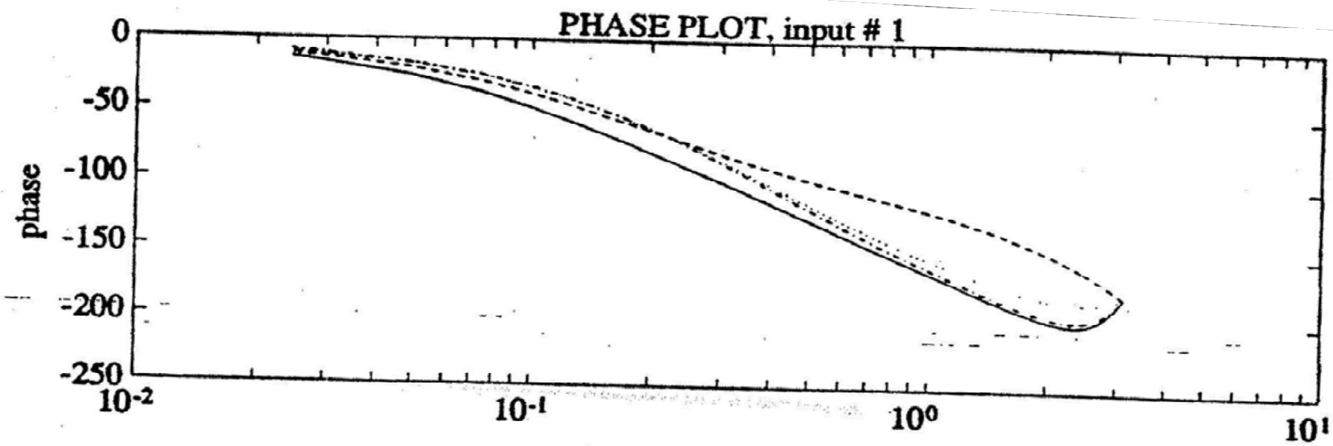
$N \to \infty \qquad \hat{\theta}_N \to \theta^*$

Results for $SNR = 1$ are better in this case than $(\cos w_o t)$.

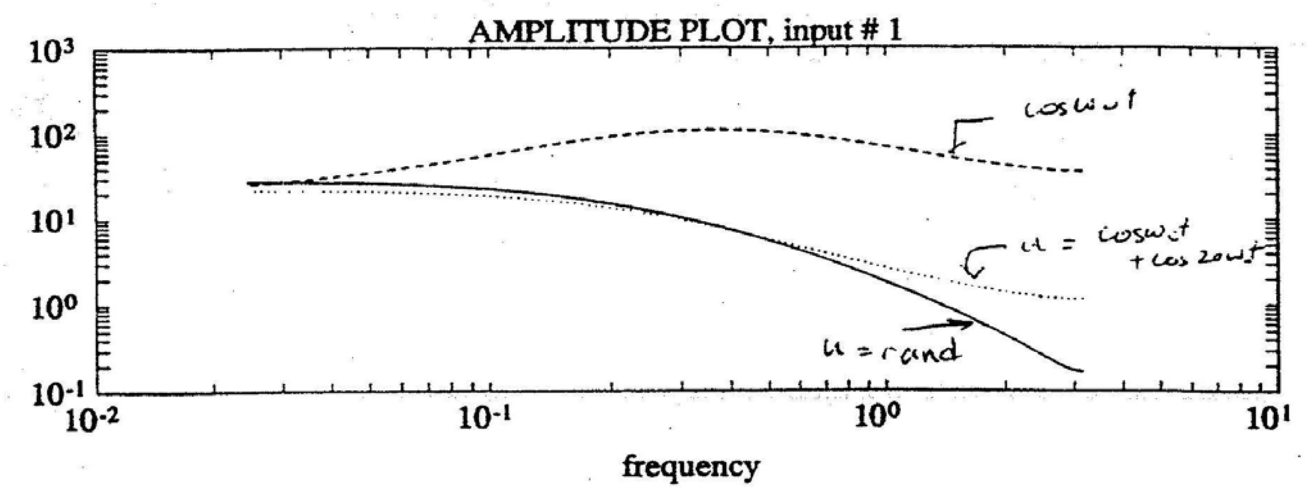


FFT of 3rd Input

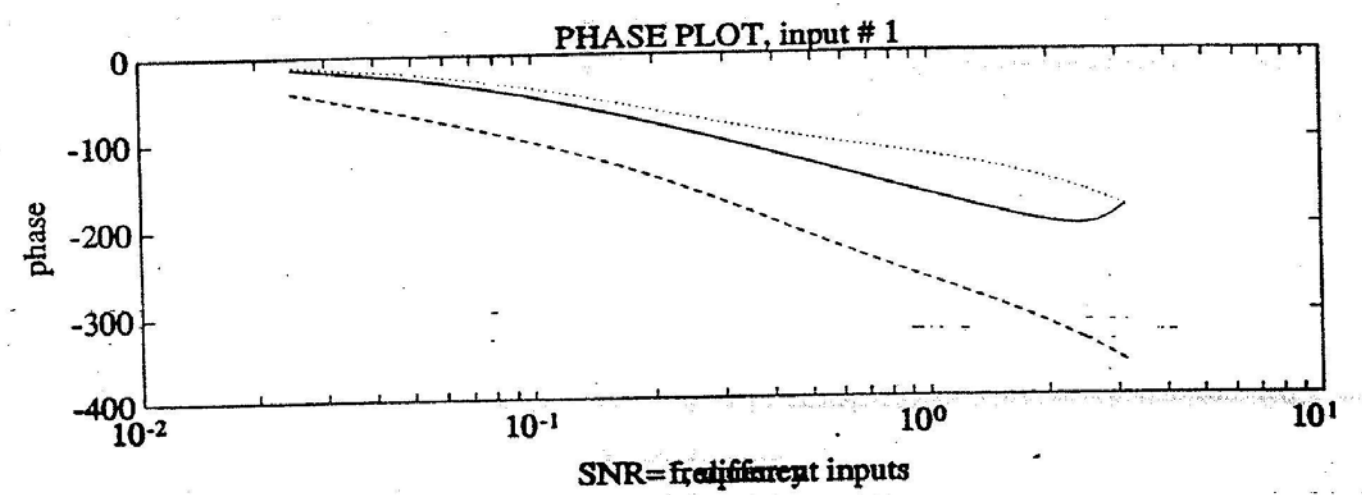




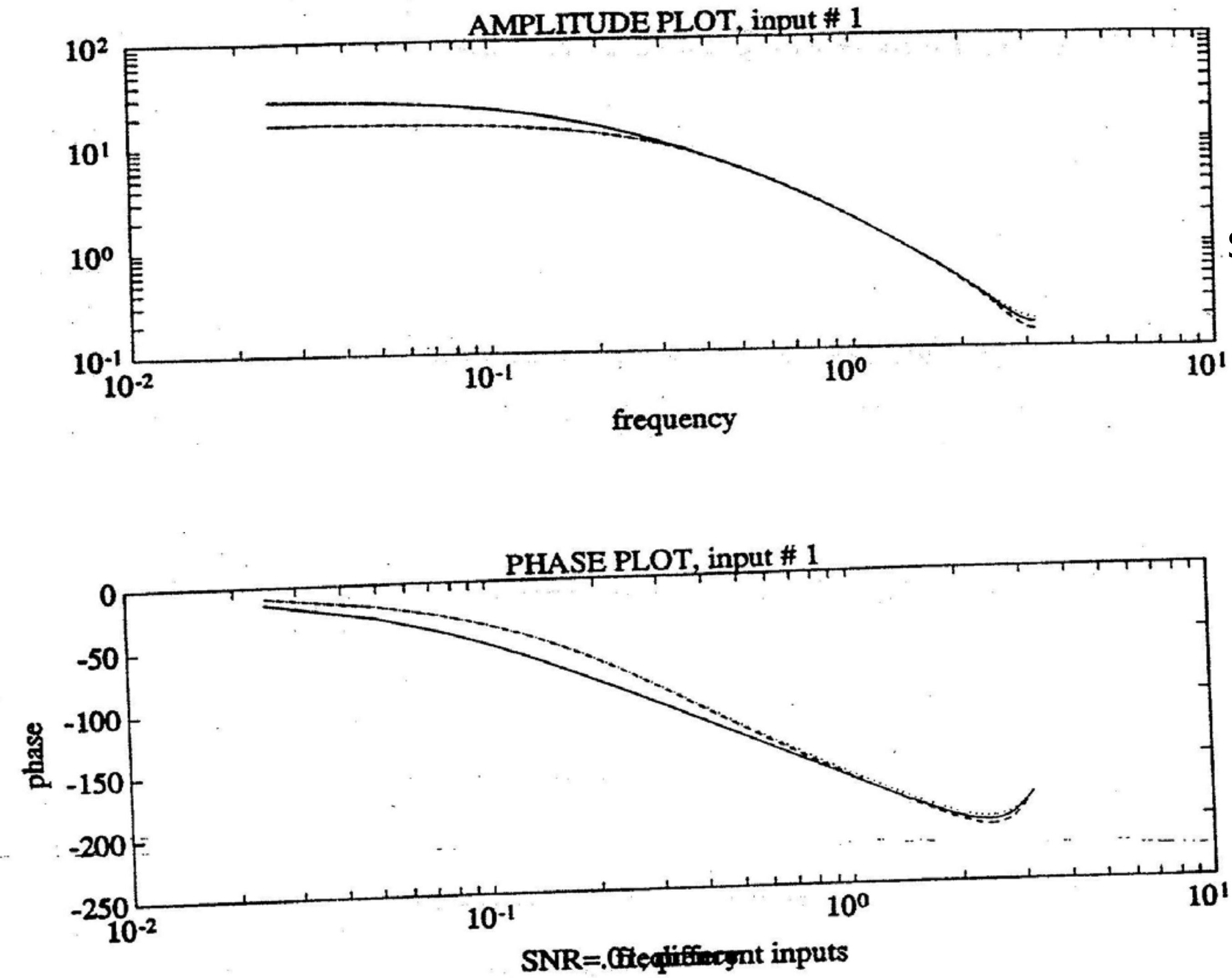
ARX (of an ARX plant)



SNR = 1



ARX (of an ARX plant)



SNR = 0.01

ARX (of an ARX plant)

Conclusions

- Accuracy of estimates depend on $\lambda\left(\bar{E}\Psi(t,\theta^*)\Psi^T(t,\theta^*)\right)^{-1}$

  If ARX, $\phi = \Psi$

- Estimate of accuracy $\simeq \hat{\lambda}_N \left(\sum_{t=1}^{N} \bar{E}\phi(t)\phi^T(t)\right)^{-1} = \hat{P}_\theta$

  λ Large $\Rightarrow \hat{P}_\theta$ is large.

  $\boldsymbol{u}$ is not rich $\Rightarrow \sum \bar{E}\phi(t)\phi^T(t)$ is singular.

  $\Rightarrow \hat{P}_\theta$ is large.

  $\boldsymbol{u}$ is rich (but close to not rich) $\Rightarrow \hat{P}_\theta$ is large.

  λ small $\Rightarrow$ better accuracy.

- Explanation (Proof of HW#3 problem 2)

$$\bar{E}\phi\phi^T(t) = \bar{E}\left[\begin{array}{cccccc} -x(t-1) & \dots & -x(t-n) & u(t-1) & \dots & u(t-n) \end{array}\right]^T \left[\begin{array}{c} \dots\dots \end{array}\right]$$

$$+ \quad \bar{E}\begin{bmatrix} v(t-1) \\ \vdots \\ v(t-n) \\ 0 \\ \vdots \\ 0 \end{bmatrix}\left[\begin{array}{cccccc} v(t-1) & \dots & v(t-n) & 0 & \dots & 0 \end{array}\right]$$

$$x = \frac{B}{A}u \qquad v = \frac{1}{A}e$$

$$= \begin{bmatrix} A_{11} & A_{12} \\ & \\ A_{21} & A_{22} \end{bmatrix} + \begin{bmatrix} \widehat{A} & 0 \\ & \\ 0 & 0 \end{bmatrix}$$

$$A_{22} = \bar{E} \begin{pmatrix} u(t-1) \\ \vdots \\ u(t-n) \end{pmatrix} \begin{pmatrix} u(t-1) & \dots & u(t-n) \end{pmatrix}$$

A_{22} is near singular $\Rightarrow$ det $\bar{E}\phi\phi^T \simeq 0$

$\Rightarrow$ Low accuracy for estimates. This can be countered by a small λ.

- How about other structures?

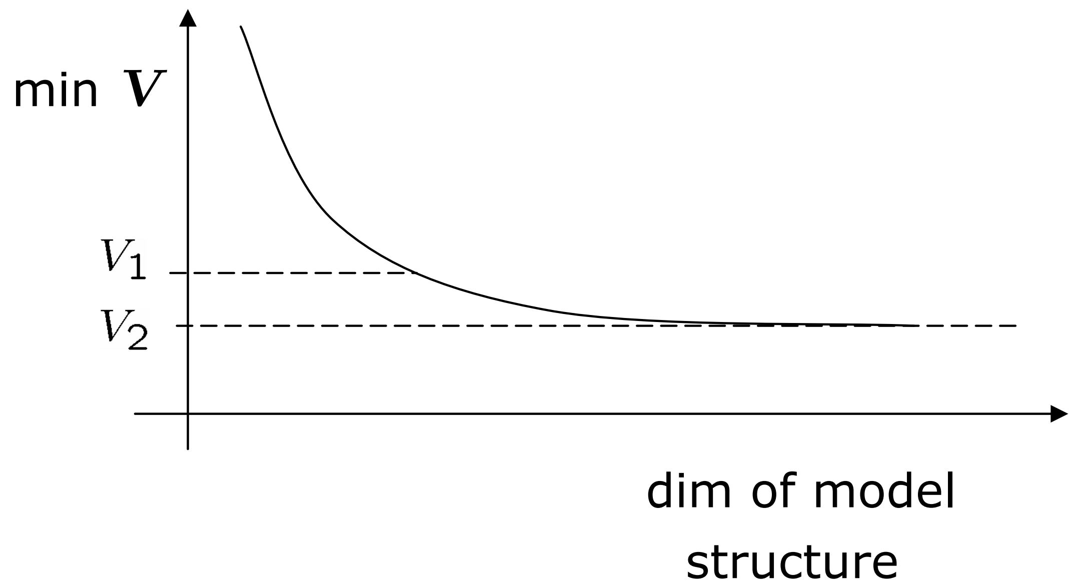
$$Ay = Bu + v$$

"same conclusions as long as v is persistent."

OE, ARMAX, ……

Comparisons of Different Model Structures

- There is a trade off between the model structure complexity and the min. error



- Akaike's Final Prediction Error criterion (FPE)

model structure

$$J(m) = \frac{1 + dm/N}{1 - dm/N} V_N\left(\hat{\theta}_N, Z^N\right)$$

$$V_N\left(\hat{\theta}_N, Z^N\right) = \min_{\hat{\theta}_N} \frac{1}{N}\sum_{t=1}^{N} \varepsilon^T \varepsilon(t|\theta)$$

- Based on minimum max likelihood. Tradeoff dm <u>vs</u> V_N

- $\widehat{\theta}_N^k = \underset{\theta \in D_{m'_R}}{\text{argmin}}\ V_N\left(\theta, Z^N\right)$

  $\bar{V}(\theta) = \lim_{N \to \infty} V_N\left(\theta, Z^N\right)$

- A natural way to evaluate a model structure m_k is by

  $$E\bar{V}\left(\widehat{\theta}_N^k\right) = J(m_k)$$

  $\widehat{\theta}_N^k$ is a random variable $\sim$ as $N(0, P_\theta)$

- Obtain estimates of both $\bar{V}(\theta)$ and $\boldsymbol{J}$
- Result

  $$J(m) = E\bar{V}\left(\widehat{\theta}_N\right) \simeq EV_N\left(\widehat{\theta}_N, Z^N\right) + \text{tr}\ \bar{V}''(\theta^*)P_N$$

  $$\simeq V_N\left(\widehat{\theta}_N, Z^N\right) + \frac{1}{N}\ \text{tr}\ \bar{V}''(\theta^*)P_\theta$$

  $$\theta^* = \text{argmin}\ \bar{V}(\theta)$$

Proof: expand $\bar{V}(\theta)$ around θ^*

$$\bar{V}\left(\widehat{\theta}_N\right) = \bar{V}(\theta^*) + \frac{1}{2}\left(\widehat{\theta}_N - \theta^*\right)^T \bar{V}''(\xi_N)\left(\widehat{\theta}_N - \theta^*\right)$$

also

$$V_N\left(\widehat{\theta}_N, Z^N\right) = \bar{V}(\theta^*, Z^N) - \frac{1}{2}\left(\widehat{\theta}_N - \theta^*\right)^T \bar{V}_N''(\bar{\xi}_N, Z^N)\left(\widehat{\theta}_N - \theta^*\right)$$

Notice:

$$E\frac{1}{2}\left(\widehat{\theta}_N - \theta^*\right)^T \bar{V}''(\xi_N)\left(\widehat{\theta}_N - \theta^*\right) = \frac{1}{2}E \text{ trace } \left\{\bar{V}''(\xi_N)\left(\widehat{\theta}_N - \theta^*\right)\left(\widehat{\theta}_N - \theta^*\right)^T\right\}$$

$$\simeq \frac{1}{2} \text{ trace } \left(\bar{V}''(\theta_o^*)P_N\right) \simeq \frac{1}{2N} \text{ trace } \bar{V}''(\theta^*)P_\theta$$

and

$$EV_N\left(\theta^*, Z^N\right) \simeq \bar{V}(\theta^*)$$

$$\Rightarrow \quad E\bar{V}\left(\widehat{\theta}_N\right) \cong \bar{V}(\theta^*) + \frac{1}{2}\ \text{trace}\ \bar{V}''(\theta^*)P_N$$

$$EV_N\left(\widehat{\theta}_N, Z^N\right) \simeq \bar{V}(\theta^*) - \frac{1}{2}\ \text{trace}\ \bar{V}''(\theta^*)P_N$$

$$J(m) = E\bar{V}\left(\widehat{\theta}_N\right) \simeq EV_N\left(\widehat{\theta}_N, Z^N\right) + \text{trace}\ \bar{V}''(\theta^*)P_N$$

$$\simeq V_N\left(\widehat{\theta}_N, Z^N\right) + \frac{1}{N}\ \text{trace}\ \bar{V}''P_\theta$$

Akaike's Information Theoretic Criterion

- Let $V_N\left(\widehat{\theta}_N, Z^N\right) = -\frac{1}{N}$ (Log likelihood function)
- Assume $\theta^* = \theta_o =$ true system
- The matrix $\bar{V}''(\theta_o)$ is invertible (identifiability).
- $\text{Cov }\widehat{\theta}_N \simeq \frac{1}{N}\left[\bar{V}''(\theta_o)\right]^{-1} = \left(E{L_N}''(\theta_o)\right)^{-1}$
- $J(m) = -\frac{1}{N}L_N\left(\theta, Z^N\right) + \text{trace }\bar{V}''(\theta_o)P_N$

$$= -\frac{1}{N}L_N\left(\theta, Z^N\right) + \frac{\text{dim }D_m}{N}^{= d\mu}$$

- Model structure determination problem

$$\left\{\widehat{\theta}_N^m, m\right\} = \underset{\substack{m \in M \\ \widehat{\theta}^n \in D_m}}{\text{argmin}} \; \frac{1}{N}\left[-L_N\left(\theta, Z^N\right) + \frac{d\mu}{N}\right]$$

- For every fixed $m, \dfrac{d\mu}{N}$ does not affect the min.

Example

- Assume that the innovations $(\varepsilon(t,\theta))$ are Gaussian with unknown variance.

- $L_N\left(\theta, Z^N\right) = -\sum_{t=1}^{N} \frac{\varepsilon^2(t,\theta)}{2\lambda} - \frac{N}{2}\log\lambda - \frac{N}{2}\log 2\pi$

- $\left(\hat{\theta}_N, \hat{\lambda}_N\right) = \underset{\theta \in D_m}{\text{argmin}} \left(-\frac{1}{N} L_N\left(\theta, Z^N\right)\right)$

$$\hat{\lambda}_N = \frac{1}{N}\sum_{t=1}^{N} \varepsilon^2(t,\hat{\theta}_N)$$

$$\hat{\theta}_N = \text{argmin} \sum_{t=1}^{N} \varepsilon^2(t,\theta)$$

- $L_N\left(\theta, Z^N\right) = -\frac{N}{2} - \frac{N}{2}\log\widehat{\lambda}_N - \frac{N}{2}\log 2\pi$

- $\widehat{m} = \text{argmin}\ \left[\frac{1}{2} + \frac{1}{2}\log\widehat{\lambda}_N^m + \frac{1}{2}\log 2\pi + \frac{dm}{N}\right]$

- Approximately minimize

$$\log\left[\frac{1}{N}\sum_{t=1}^{N}\varepsilon^2(t, \widehat{\theta}_N)\right] + \frac{2dm}{N}$$

Akaike's Final Prediction Error Criterion

- $J(m) \simeq V_N\left(\widehat{\theta}_N, Z^N\right) + \frac{1}{N}\ \text{trace}\ \bar{V}''(\theta^*)P_\theta$
- Let $V_N\left(\widehat{\theta}_N, Z^N\right) = \frac{1}{2N}\sum_{t=1}^{N}\varepsilon^2(t,\theta) \qquad \theta_o \in m$

$$\therefore \quad P_\theta = \lambda_o\left(\bar{V}''(\theta_o)\right)^{-1} \qquad P_N \simeq \frac{1}{N}P_\theta$$

$$\lambda_o = 2\bar{V}(\theta_o) = \left(E{e_o}^2(t) = E\varepsilon^2(t,\theta_o)\right)$$

$$\therefore \quad J(m) = V_N\left(\widehat{\theta}_N, Z^N\right) + \text{trace}\ \frac{1}{N}\left(\bar{V}''(\theta_o)P_\theta\right)$$

$$= V_N\left(\widehat{\theta}_N, Z^N\right) + \lambda_o\frac{dm}{N}$$

- $V_N\left(\hat{\theta}_N, Z^N\right) \simeq \bar{V}(\theta_o) - \frac{1}{2}\ \text{trace}\ \bar{V}''(\theta_o)P_N$

$$\simeq \frac{\lambda_o}{2} - \frac{1}{2}\frac{dm}{N}\lambda_o$$

estimate of λ_o
$$\hat{\lambda}_N = \left[\frac{2V_N\left(\hat{\theta}, Z^N\right)}{N - dm}\right]\frac{N}{1}$$

$$\boxed{\begin{aligned} J(m) &= \frac{1 + dm/N}{1 - dm/N} V_N\left(\hat{\theta}_N, Z^N\right) \\ &= \frac{1 + dm/N}{1 - dm/N}\frac{1}{N}\sum_{t=1}^{N}\frac{1}{2}\varepsilon^2(t, \hat{\theta}_N) \end{aligned}}$$

(FPE)

Example 2

- δ is unknown.
- 3 experiments

$$u = PRBS$$

$$u = \cos \omega_o t \qquad w_o = 2\pi/1000$$

$$u = \cos 20 w_o t + \cos w_o t.$$

- Consider 2 - model structures

$$m_1 : \quad Ay = Bu + e$$

$$m_2 : y = \frac{B}{F} u$$

Experiment 1

- u = rand sequence
- $\widehat{R}_s \cong \bar{E}\phi_s\phi_s^T(t)$

$$= \frac{1}{N}\sum_{t=1}^{N} \phi_s(t)\phi_s^T(t)$$

$s = 2 \quad \text{det} = 0.062$

$s = 3 \quad \text{det} = 3 \times 10^{-7} \simeq 0$

$\Rightarrow$ system has dim = 2

- Estimated parameters

	a_1	a_2	b_1	b_2	$\bar{E}\varepsilon\varepsilon^T$
ARX	-1.4	0.49	1	0.5	0
OE	-1.4	0.49	1	0.5	0

- Try different structures (n_a, n_b, n_k)

$$(1 \quad 1 \quad 1) \rightarrow 0.1022$$

$$(2 \quad 2 \quad 1) \rightarrow 0$$

$$(3 \quad 3 \quad 2) \rightarrow 0.0710$$

$$(3 \quad 4 \quad 1) \rightarrow 0$$

AIC or FPE $\Rightarrow (2 \quad 2 \quad 1)$ is the best choice.

Experiment 2

- $u = \cos w_o t$ $\quad\quad$ $\boldsymbol{u}$ is p.e of order 2

- $\hat{R}_s \simeq \frac{1}{N} \sum_{t=1}^{N} \phi_s(t) \phi_s^T$

$$\det\left(\hat{R}_2\right) = 2.66 \times 10^{-7}$$

$$\det\left(\hat{R}_3\right) = 3.7 \times 10^{-33}$$

$$\det\left(\hat{R}_1\right) = 8.77$$

<u>ARX</u>: $n \geq 1$ it may be $n = 2$ or 3 ......

<u>OE</u>: $n \geq 1$ " " " " ......

Structure	ARX Parameters	V_N	OE Parameters	V_N
(1, 1, 1)	(-0.8627, 2.3182)	(0.034)	(-0.85, 2.51)	0.214
(2, 2, 1)	(-1.4, 0.49, 1, 0.5)	0	(-1.4, 0.49, 1, 0.5)	0
(3, 3, 1)	(- * * * * ......)	170	(* * * * ......)	0.3317
⋮				

Loss of identifiability.

Clearly $(2 \quad 2 \quad 1)$ is the preferred model structure

Remark: Both experiments were generated from the model

$$\delta: \qquad y(t) = G_o u \qquad G_o = \frac{q^{-1}\left(1 + 0.5q^{-1}\right)}{1 - 1.4q^{-1} + 0.49q^{-2}}$$

Experiment 3

- $u = \cos w_o t$
- $\det\left(\hat{R}_2\right) = 2.65 \times 10^{-7} \qquad n \geq 1$

Structure	ARX Parameters	V_N	OE Parameters	V_N
(2 2 1)	(-1.4004, 0.4903, 0.98, 0.51)	3.4×10^{-5}	(-1.40, 0.49, 0.95, 0.54)	9.2×10^{-6}
(3 3 1)	singular		singular	

OE model is preferred.

- $u = \cos 20 w_o t + \cos w_o t$
- $\det\left(\hat{R}_2\right) = 0.0692 \qquad \det\left(\hat{R}_3\right) = 3.9 \times 10^{-11}$

Structure	ARX Parameters	V_N	OE Parameters	V_N
(2 2 1)	(-1.4, 0.49, 1.0031, 0.4944)	1.6×10^{-5}	(1.399, 0.489, 0.99, 0.50)	1.3×10^{-5}
(3 3 1)	(* * * ......)	1.3×10^{-5}	(* * * ......)	4.2×10^{-5}

Num. errors

AIC for ARX $\Rightarrow (3 \quad 3 \quad 1)$ {In fact $(4 \quad 4 \quad 1)$ does better}.

AIC for ARX and OE $\Rightarrow (2 \quad 2 \quad 1)$ OE structure.

Remark: Data generated by

$$\delta : \qquad y(t) = G_o u + 0.01 e$$

Validation

- Use different sets of data to validate the model structure and the estimated model.

- You can obtain different estimates using the data and then <u>average them</u>. OR you can construct <u>new</u> input-output pairs and re-estimate.

Conclusions

- Criterion contains a penalty function for the dimension of the system.

- "AIC" is one way of doing that. (FPE) is an estimate of the AIC with a quadratic objective.

- "AIC" has connections with information theory observation

$$Z^t \longrightarrow \text{ assumed PDF } f_m\left(t, Z^t\right)$$

$$\text{true PDF } f_o\left(t, Z^t\right)$$

Entropy of f_o w. r. to $f_m = \delta(f_o, f_m) = -I(f_o, f_m)$

$$I(f_o, f_m) = \int f_o\left(t, x^t\right) \log \frac{f_o\left(t, x^t\right)}{f_m\left(t, x^t\right)} dx^t.$$

over the
observation
= information distance

$$\hat{\theta}_N = \underset{\theta}{\text{argmin}}\ I\left(f_o, f_m\left(\theta, Z^N, N\right)\right)$$

- "AIC" is the average information distance

$$E_{\hat{\theta}_N} I(f_o, f)$$

after some simplification

$$\hat{\theta}_{AIC} = \underset{\theta}{\text{argmin}}\ \left\{\frac{1}{N}\sum_{t=1}^{N} l(\varepsilon(t,\theta), t, \theta) + \frac{\text{dim }\theta}{N}\right\}$$

- Careful about Numerical errors!?