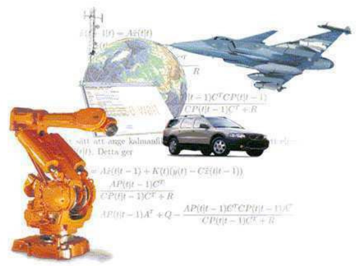
# Sysid Course VT1 2016
# Linear Models – Special Issues

## Chapters 6 and 7 in Text Book



**Lennart Ljung**

Automatic Control, ISY, Linköpings Universitet



## Outline



- Frequency Domain Data: Parametric and Nonparametric Fitting
- The Instrumental Variable Method
- Subspace Techniques
- Regularization



## Goal



Goal: Estimate a linear model in discrete or continuous time with or without an additive noise model.

$$y(t) = G(\sigma)u(t) + v(t)$$

$\sigma$ is differentiation operator $p$ or shift operator $q$.
The corresponding frequency response function (FRF) is $G(i\omega)$ or $G(e^{i\omega})$. Estimating a linear system is the same as estimating its FRF-curve.



## Recall: The Frequency Response Function, FRF

$\rightarrow$ 170



A linear system is characterized by its transfer function $G(s)$ (the Laplace transform of its impulse response).
Evaluated on the imaginary axis, this gives the FRF $G(i\omega)$, which describes the response to sinusoidal inputs:

$$u(t) = A\cos(\omega t), \quad y(t) = A_1\cos(\omega t + \phi)$$
$$A_1 = |G(i\omega)|A, \quad \phi = \arg G(i\omega)$$

This could be a way of determining $G$ (frequency analysis).
Discrete time: $G(z)$, $z$-transform, unit circle, $G(e^{i\omega})$
All Frequencies at the same time:

$$Y(i\omega) = G(i\omega)U(i\omega) + \text{transient}$$

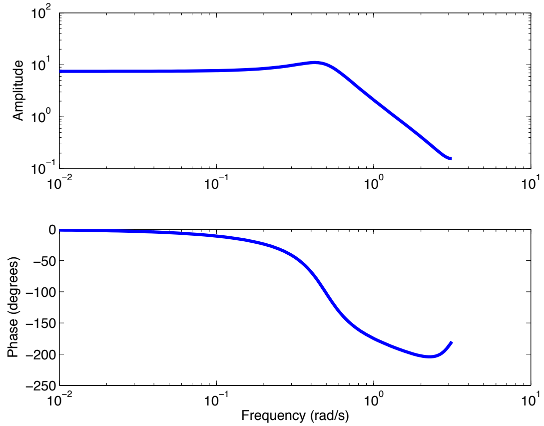
$Y$ and $U$ are the Fourier transforms of the output and input.
.

## The Bode Plot

$\rightarrow$ 28



$|G(i\omega)|$ and arg $G(i\omega)$ vs $\omega$






## Data from Dynamic Systems: Input-Output Data

$\rightarrow$ 228



- Discrete time
  - Time-domain: $\{u(1), y(1), u(2), y(2), \ldots, u(N), y(N)\}$
  - Frequency-domain $\{U_N(e^{i\omega_1}), Y_N(e^{i\omega_1}), \ldots, U_N(e^{i\omega_N}), Y_N(e^{i\omega_N})\}$ DFT-grid: $\omega_k = 2\pi k/N$

$$U_N(z) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} u(k) z^{-k}$$

- Continuous time
  - Frequency-domain $\{U_N(i\omega_1), Y_N(i\omega_1), \ldots, U_N(i\omega_N), Y_N(i\omega_N)\}$

$$U_N(s) = \frac{1}{\sqrt{N}} \int_0^N u(t) e^{-st} dt$$

(Band limited, periodic data)



## Frequency Response Data (FRD)

$\rightarrow$ 173



- From frequency analyzers or computed/estimated using FFT techniques
  - $\hat{\hat{G}}(i\omega_k)$ or $\hat{\hat{G}}(e^{i\omega_k}), k = 1, 2, \ldots, N$
  - Possibly with uncertainty measures $W(i\omega_k)$
  - Simple estimate, ETFE:

$$\hat{\hat{G}}_N(i\omega) = \frac{Y_N(i\omega)}{U_N(i\omega)} \qquad \text{Variance: } W(i\omega) = \frac{\Phi_v(\omega)}{|U_N(i\omega)|^2}$$

where $\Phi_v(\omega)$ is the spectrum of the output disturbance

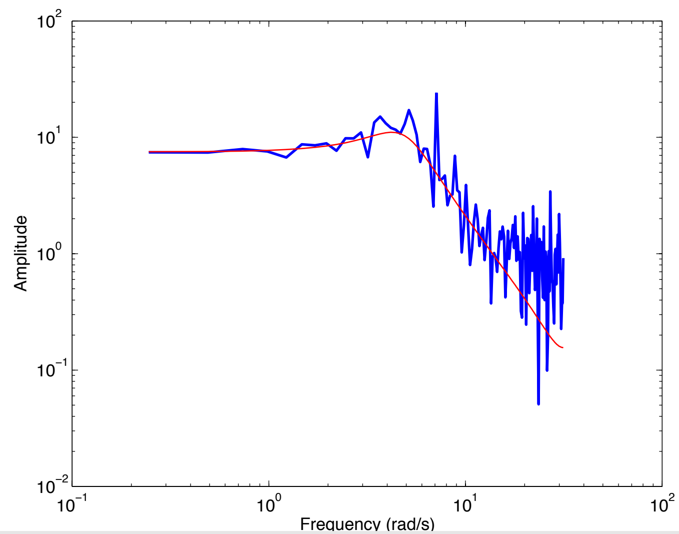
  - Other estimates (spectral analysis): smoothed versions of ETFE (more later)



## ETFE: Empirical Transfer Function Estimate

$\rightarrow$ 173





## Recall: Linear Parametric Models



From Introductory notes: Any linear model. Take the Fourier Transforms of the signals:

$$y(t) = G(q, \theta)u(t) + H(q, \theta)e(t)$$
$$FT: Y(e^{i\omega}) = G(e^{i\omega}, \theta)U(e^{i\omega}) + H(e^{i\omega}, \theta)E(e^{i\omega})$$

Exampels of parametrizations of $G(q, \theta), \ldots$

$$G(q, \theta) = \frac{B(q)}{F(q)}; \quad H(q, \theta) = \frac{C(q)}{D(q)}$$
$$y(t) = \frac{B(q)}{A(q)}u(t) + \frac{1}{A(q)}e(t) \text{ or}$$
$$G(q, \theta) = C(\theta)(qI - A(\theta))^{-1}B(\theta).$$
$$H(q, \theta) = C(\theta)(qI - A(\theta))^{-1}K(\theta) + I$$



## Prediction Error Fit to Parametric Models

→ 230



- With a linear model, and a quadratic prediction error loss $(\ell(\varepsilon)) = \varepsilon^2)$ we can apply Parseval's relation to the criterion function $\sum \varepsilon^2(t)$, and with $Y_N$ and $U_N$ being the DFTs:

$$V_N(\theta) = \sum_{k=1}^{M} |Y(e^{i\omega_k}) - G(e^{i\omega_k}, \theta)U_N(e^{i\omega_k})|^2 / |H(e^{i\omega_k}, \theta)|^2 \text{ or}$$

$$V_N(\theta) = \sum_{k=1}^{M} \left| \frac{Y_N(e^{i\omega_k})}{U_N(e^{i\omega_k})} - G(e^{i\omega_k}, \theta) \right|^2 \cdot \left| \frac{U_N(e^{i\omega_k})}{H(e^{i\omega_k}, \theta)} \right|^2$$

- Formal and intuitive interpretation ...

$$\varepsilon = \frac{1}{H}(y - Gu)$$

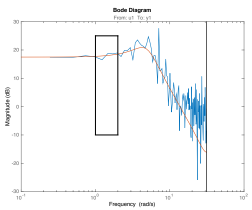


## Non-Parametric Methods: Smoothing the ETFE

→ 178



Intuitively, think of sliding a window of width $BW$ along the ETFE and average what you see in the window:




The width $BW$ will affect the variance of the smoothed estimate [Large $BW \Rightarrow$ many values to average $\Rightarrow$ small variance.] and the bias and frequency resolution [small $BW \Rightarrow$ small bias and near-by peaks are distinguishable.]



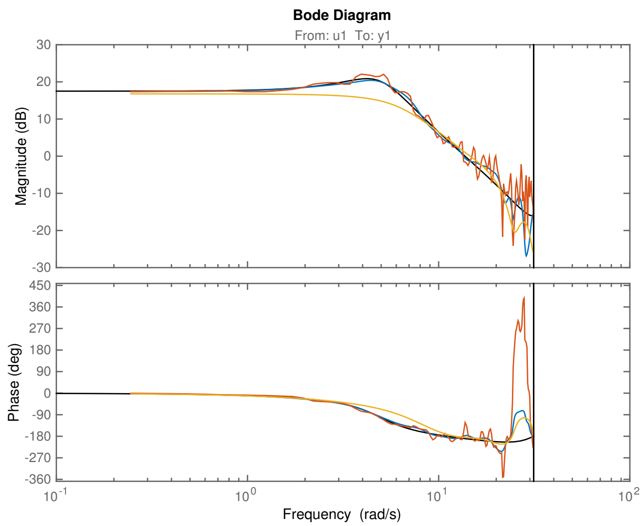
## Spectral Analysis: Blackman-Tukey

→ 181



The smoothing of the EFTE can also be done on time domain data.

1. Form the covariance functions $\hat{R}_u(\tau) = \frac{1}{N}\sum_{t=1}^{N} u(t)u(t-\tau)$ (and similar for $\hat{R}_{yu}(\tau)$)
2. Form the weighted Fourier transforms $\hat{\Phi}_u(\omega) = \sum_{\tau=-\infty}^{\infty} w_\gamma(\tau)\hat{R}_u(\tau)$ Same for $\hat{\Phi}_{yu}(\omega)$
3. Form the estimate $\hat{G}(e^{i\omega}) = \frac{\hat{\Phi}_{yu}(\omega)}{\hat{\Phi}_u(\omega)}$
4. the (time window) $w_\gamma(\tau)$ is the inverse FT of the frequency window that was slided along the ETFE on the previous slide. Often $w_\gamma(\tau) = 0$ for $|\tau| > \gamma$. Due to the time/frequency links $\gamma \sim 1/BW$.
5. So large $\gamma$ (small $BW$) means good frequency resolution and large variance and v.v.




Black: True
Blue: $\gamma = 30$
Red: $\gamma = 100$
Yellow: $\gamma = 10$
BW $\sim 2\pi/\gamma$











Consider the ARX-model

$$A(q)y(t) = B(q)u(t) + w(t)$$

or $y(t) = \varphi^T(t)\theta_0 + w(t)$

$$\varphi^T(t) = \begin{bmatrix} -y(t-1) & \dots & -y(t-n) & u(t-1) & \dots & u(t-n) \end{bmatrix}$$

$$\theta_0 = \begin{bmatrix} a_1 & \dots & a_n & b_1 & \dots b_n \end{bmatrix}^T$$

$$\hat{\theta} = [\sum \varphi(t)\varphi^T(t)]^{-1} \sum \varphi(t)y(t)$$

$$\hat{\theta} = \theta_0 + [\sum \varphi(t)\varphi^T(t)]^{-1} \sum \varphi(t)w(t)$$

If $w$ is not white, $\varphi(t)$ and $w(t)$ are correlated and $\theta$ will be biased!





Consider the linear regression $\hat{y}(t|\theta) = \varphi^T(t)\theta$ (This could be an ARX-model, but could also be something else)

Suppose that the data is generated by

$$y(t) = \varphi^T(t)\theta_0 + w(t)$$

for some noise sequence $w(t)$.

Choose a sequence of vectors – the instruments – $\zeta(t)$ (of the same dimension as $\varphi$). Multiply it with the equation above and sum over $t$:

$$\frac{1}{N}\sum \zeta(t)y(t) = \frac{1}{N}\sum \zeta(t)\varphi^T(t)\theta + \frac{1}{N}\sum \zeta(t)w(t)$$

which suggests the estimate

$$\hat{\theta}_N = [\frac{1}{N}\sum_{t=1}^{N} \zeta(t)\varphi^T(t)]^{-1}\frac{1}{N}\sum_{t=1}^{N} \zeta(t)y(t)$$

(Note that $\zeta(t) = \varphi(t)$ gives the least squares method!)

## Choice of Instruments

$\longrightarrow$ 225



Note that

$$\hat{\theta}_N = \theta_0 + [\frac{1}{N}\sum_{t=1}^{N}\zeta(t)\varphi^T(t)]^{-1}\frac{1}{N}\sum_{t=1}^{N}\zeta(t)w(t)$$

$\zeta(t)$ are the *instruments*. The requirements on them are

1. $\zeta(t)$ and $w(t)$ be *uncorrelated*
2. $\zeta(t)$ and $\varphi(t)$ be *correlated* so that the indicated inverse in $\hat{\theta}_N$ exists.

Under these assumptions $\hat{\theta}_N$ will converge to the true value of the parameters as the number of data tends to infinity.



## How to achieve this for ARX models?



Basic choice: Choose $\zeta(t)$ as the "noise free" counterpart of $\varphi(t)$
More specifically: Let $N(q)$ and $M(q)$ be two filters and define $x(t)$ from the input sequence as

$$N(q)x(t) = M(q)u(t)$$

and take

$$\zeta(t) = \begin{bmatrix}-x(t-1) & \cdots & -x(t-na) & u(t-1) & \cdots & u(t-nb)\end{bmatrix}^T$$

It can be shown that for "almost all" choices of filters $M$ and $N$ (of orders at least as large as the model) this will satisfy the two requirements on the previous slide



## Optimal Choices of Instruments



What are the "best choices" of filters $N$ and $M$?
To answer that we must consider a variant of IV that allows prefiltering:

$$\hat{\theta}_N = [\frac{1}{N}\sum_{t=1}^{N}\zeta(t)\varphi_F^T(t)]^{-1}\frac{1}{N}\sum_{t=1}^{N}\zeta(t)y_F(t)$$

$$\varphi_F(t) = L(q)\varphi(t)$$

$$y_F(t) = L(q)y(t)$$

A general, but somewhat complicated, expression for how the variance of $\hat{\theta}_N$ depends on $L$, $M$ and $N$ can be given. The choices that minimize this covariance matrix depend on the true system.



## Optimal Choices of Instruments, cnt'd



Suppose that the true system is given by

$$y(t) = G_0(q)u(t) + H_0(q)e(t)$$

$$G_0(q) = \frac{B_0(q)}{A_0(q)}$$

where $H_0$ is known and $G$ to be estimated. The optimal choices of instruments – in the sense that the variance of the estimates is minimized – is then obtained for

- $N(q) = A_0(q)$
- $M(q) = B_0(q)$
- $L(q) = \frac{1}{A_0(q)H_0(q)}$

## The IV4 Algorithm



In practice, $G_0$ and $H_0$ are not known. A feasible way of choosing almost optimal instruments is then the following 4-step method (*iv4*):

1. Estimate $\hat{A}_1$ and $\hat{B}_1$ using LS
2. Use IV with $L = 1, N = \hat{A}_1$ and $M = \hat{B}_1$. This gives $\hat{A}_2$ and $\hat{B}_2$.
3. Calculate the residuals $w(t) = \hat{A}_2(q)y(t) - \hat{B}_2(q)u(t)$ and fit a filter $L$ to the AR-model $L(q)w(t) = e(t)$ using LS. This gives $\hat{L}(q)$
4. Use IV with $L = \hat{L}, N = \hat{A}_2$ and $M = \hat{B}_2$. This gives the final estimates.



## Summary: The IV Method



- Good method to quickly get the dynamics of a system
- Be careful when data have been collected in closed loop
- Does not provide a noise model
- Good alternative to OE-structures.



## Outline



- Frequency Domain Data: Parametric and Nonparametric Fitting
- The Instrumental Variable Method
- Subspace Techniques
- Regularization



## Subspace Methods

$\longrightarrow$ 208



$$x(t+1) = Ax(t) + Bu(t) + w(t)$$
$$y(t) = Cx(t) + Du(t) + v(t)$$

Estimate the matrices $A, B, C, D$.
Suppose, for a second, that the states $x(t)$ were known. Then the above expression is a linear regression: Let

$$Y(t) = \begin{bmatrix} x(t+1) \\ y(t) \end{bmatrix}$$
$$\Phi(t) = \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}$$

## The Linear Regression

Then

$$Y(t) = \Theta\Phi + \nu(t)$$

with

$$\Theta = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$
$$\nu = \begin{bmatrix} w(t) \\ v(t) \end{bmatrix}$$

All matrices of interest, including the covariance matrix of $\nu$ could then be estimated using the Least Squares method. With the covariance matrix of $\nu$, the optimal Kalman gain could then be computed.

.





## How to Find the States and Conceptual Algorithm

→ 210

Fact: All (interesting) states can be found as linear combinations of the $k$-step ahead predictors $\hat{y}(t+k|t)$, $k = 1, \ldots, n$ (the predicted value of $y(t+k)$ based on input-output data up to time $t$. No prediction of the effect of inputs after time $t$.)

So estimate these $k$-step ahead predictors using ARX-models, and determine from these the good linear combinations to form the states $x$.

Use these $x$ to form the linear regression to estimate $A, B, C, D$.





## Subspace Algorithms: Essential Features

- State space basis selected automatically
- Form sample covariances of $y$ and $u$: One SVD and one QR-step
- No iterations
- Need to select auxiliary variables: (essentially the ARX orders for which the predictors – state candidates – are estimated)
- Quality properties not fully understood





## More Formal Calculations, 1/5

$$y(t+k) = \sum_{j=-\infty}^{t+k} h^u_{t+k-j} u(j) + h^e_{t+k-j} e(j)$$
$$\hat{y}(t+k|t) = \sum_{j=-\infty}^{t} h^u_{t+k-j} u(j) + h^e_{t+k-j} e(j)$$

Let

$$Y^x(t) = \begin{bmatrix} \hat{y}(t+1|t) \\ \vdots \\ \hat{y}(t+n|t) \end{bmatrix}$$

.

# More Formal Calculations, 2/5



So, all (Kalman) states $x(t)$, in any state-space representation can be written as linear combinations of $Y^x(t)$:

$$x(t) = LY^x(t)$$

for some $L$. The (minimal) order of the state-space representation is the rank of $Y^x(t), t = 1, \ldots,$.
So, with $Y^x(t), t = 1, \ldots, N$ given, pick $L$, so that $x(t)$ becomes well conditioned. This includes the choice of dimension of $x$.
Typically, apply SVD to

$$Y^N = \begin{bmatrix} Y^x(1) & Y^x(2) & Y^x(N) \end{bmatrix}$$

Once $x(t), t = 1, \ldots, N$ have been determined, proceed as above to find the state-space matrices.



# More Formal Calculations, 3/5



How to estimate the predictors:

$$y(t+k) = \sum_{j=-\infty}^{t+k} h^u_{t+k-j} u(j) + h^e_{t+k-j} e(j) (*)$$

$$\hat{y}(t+k|t) = \sum_{j=-\infty}^{t} h^u_{t+k-j} u(j) + h^e_{t+k-j} e(j)$$

$e(t)$ and $y(t)$ have an invertible relationship.

$$y(t+k) = \sum_{j=-\infty}^{t+k-1} \tilde{h}^u_{t+k-j} u(j) + \tilde{h}^e_{t+k-j} y(j) + e(t+k) + \tilde{h}^u_0 u(t+k) (**)$$

so replace $e(j)$ in (*) by $y$ and $u$ from (**): . . .



# More Formal Calculations, 4/5



$$y(t+k) = \sum_{j=-\infty}^{t+k} \tilde{h}^u_{t+k-j} u(j) + \sum_{j=-\infty}^{t} \tilde{h}^e_{t+k-j} y(j) + \sum_{j=t+1}^{t+k} h^e_{t+k-j} e(j)$$

$$\hat{y}(t+k|t) = \sum_{j=-\infty}^{t} \tilde{h}^u_{t+k-j} u(j) + \tilde{h}^e_{t+k-j} y(j)$$

Now, truncate the first equation at $j = t - n_1$ rather than at $j = -\infty$, and estimate $\tilde{h}$ using the least squares method. Use these estimates in the second equation to estimate $\hat{y}$. The value $n_1$ corresponds to the "auxiliary order". All of this can be done numerically efficient by projections.

.



# The Subspace Method Algorithm, 5/5



The essence of the subspace methods is as follows

1. Select $n$ and $n_1$ and estimate $Y^x(t), t = 1, \ldots, N$.
2. Determine a good choice of $L$ in $x(t) = LY^x(t)$ (including dimension) using SVD or similar decomposition
3. Possibly determine $n$ by visual inspection of the singular values in the above expression.
4. Estimate $A, B, C$ and $D$ by least squares in the state-space model, treating $x(t)$ as a measured sequence.
5. Use the covariance matrix of $\nu$ to compute the Kalman Filter gain $K$.

# Outline



- Frequency Domain Data: Parametric and Nonparametric Fitting
- The Instrumental Variable Method
- Subspace Techniques
- Regularization ( → 221 , New stuff!)



# Recall: Bias – Variance Trade Off

→ 492



Any estimated model is incorrect. The errors have two sources:

- Bias: The model structure is not flexible enough to contain a correct description of the system.
- Variance: The disturbances on the measurements affect the model estimate, and cause variations when the experiment is repeated, even with the same input.

Mean Square Error (MSE) = $|\text{Bias}|^2$ + Variance.
When model flexibility ↑,Bias ↓ and Variance ↑.
To minimize MSE is a good trade-off in flexibility.
In state-of-the-art Identification, this flexibility trade-off is governed primarily by model order. May need a more powerful tuning instrument for bias–variance trade-off.

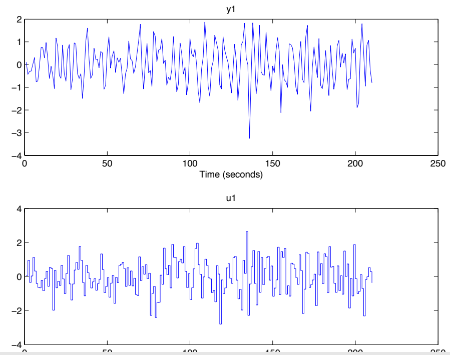


# An Example



Equipped with the tools from the previous lecture, let us now test some data `z` (selected but not untypical). The example uses complex dynamics and few (210) data, so this is a case where asymptotic properties are not prevalent. Find the Impulse Response (IR)!

```
plot(z)
```





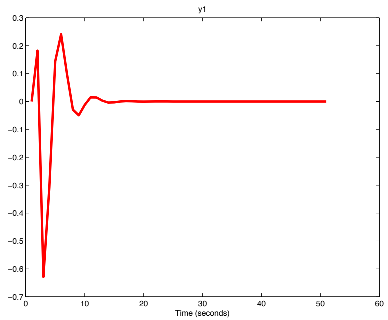

# Estimate a Model: State-of-the-Art

→ 221



We will try the state-of-the art approach: Estimate SS models of different orders. Determine the order by the AIC criterion.

```
 for k=1:30
   m{k}= ssest(z,k);
end
(dum,n) = min(aic(m{:}));
mss = m{n};
impulse(mss)
```




Is this a good model? Preview: This IR has a fit of **79.42%**
But, we can do better! Another choice of model order gives a fit of **82.95 %** . I will also show an estimate with **a 83.55%** fit.

## Linear Black-Box Models: Fundamental Role of ARX

→ 336



Recall: ARX: $A(q)y(t) = B(q)u(t) + e(t)$

**ARX can Approximate Any Linear System**

Arbitrary Linear System: $y(t) = G_0(q)u(t) + H_0(q)e(t)$

ARX model order $n, m$ : $A_n(q)y(t) = B_m(q)u(t) + e(t)$

$\hat{y} = (1 - A(q))y(t) + B(q)u(t)$ – General linear predictor!

as $N >> n, m \to \infty$

$[\hat{A}_n(q)]^{-1}\hat{B}_m(q) \to G_0(q), \; [\hat{A}_n(q)]^{-1} \to H_0(q)$

**The ARX-model Is a Linear Regression**

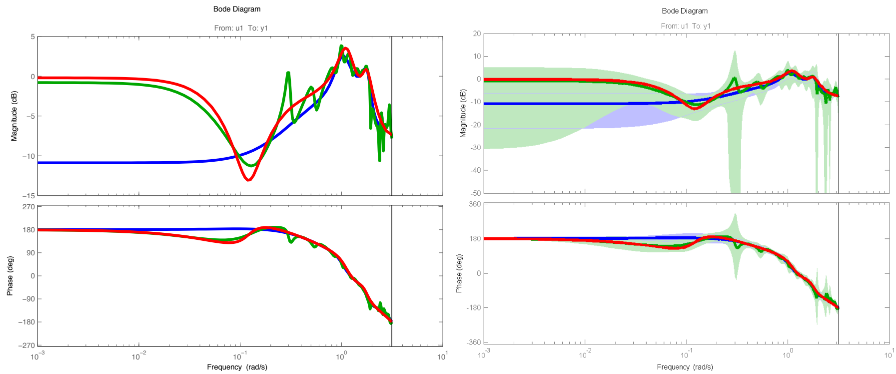
Note that the ARX-model is estimated as a linear regression
$Y = \Phi\theta + E$, ($\Phi$ containing lagged $y, u$ and $\theta$ containing $a, b$)
**A convex estimation problem.**



## Test on Our Data



Estimate ARX-model of order 10 and 30: Bode plots of models together with true system:




Order 10. Order 30. True. The high order model picks up the true curves better, but seem more "shaky". Look at Uncertainty regions!



## How to Curb Variance?



The ARX approximation property is valuable, but high orders come with high variance.
Can we curb the flexibility that causes high variance other than by lower order? Regularization



## Model Structures and Regularization

→ 221



**Curb the Model's Flexibility!**

$$V_N(\theta) = \sum_{t=1}^{N} |\varepsilon(t,\theta)|^2 + \lambda(\theta - \theta^*)^T R(\theta - \theta^*)$$

## Bayesian View

$\rightarrow 230$



**The regularized criterion**

$V_N(\theta) = \sum_{t=1}^{N} |\varepsilon(t,\theta)|^2 + \lambda(\theta - \theta^*)^T R(\theta - \theta^*)$

**Bayesian interpretation**

$\theta$ is a random vector which *a prior* (Gaussian) distribution with mean $\theta^*$ and covariance matrix $(\lambda R)^{-1}$

That means that with the regularized estimate $\hat{\theta}_N = \arg\min V_N(\theta)$ is the *Maximum A Posteriori* (MAP) Estimate.



## Linear Regression – Regularization



**The regularized criterion**

$V_N(\theta) = \sum_{t=1}^{N} |\varepsilon(t,\theta)|^2 + \lambda(\theta - \theta^*)^T R(\theta - \theta^*), \quad \lambda R = \Pi^{-1}$

Regularization for a linear regression ($\theta^* = 0$) (Recall that ARX is a linear regression.)

$$Y = \Phi\theta + E$$
$$\hat{\theta}_N = \arg\min |Y - \Phi\theta|^2 + \theta^T \Pi^{-1} \theta$$

$\Pi$ is the Regularization Matrix ( = the prior covariance matrix). Still quadratic in $\theta$: The estimate will be

$$\hat{\theta}^R = (R_N + \Pi^{-1})^{-1} R_N \hat{\theta}^{LS} \quad R_N = \Phi\Phi^T$$

How to choose $\Pi$? How good is it? : Classical (frequentist) analysis next slide



## Frequentist analysis of Regularized LSE



Assume true parameter $\theta_0 \quad Y = \Phi\theta_0 + E: \quad \mathcal{E}EE^T = I.$
BIAS:

$$\mathcal{E}\hat{\theta}^R - \theta_0 = (R_N + \Pi^{-1})^{-1}\Pi^{-1}\theta_0$$

MSE:

$$\mathcal{E}[(\hat{\theta}^R - \theta_0)(\hat{\theta}^R - \theta_0)^T] = (R_N + \Pi^{-1})^{-1} \times$$
$$(R_N + \Pi^{-1}\theta_0\theta_0^T\Pi^{-1})(R_N + \Pi^{-1})^{-1};$$

No regularization ($\Pi^{-1} = 0$): Unbiased and MSE $= R_N^{-1}$ (Cramér-Rao bound)
Best MSE?: Minimized by $\Pi = \theta_0\theta_0^T$: MSE = $(R_N + \Pi^{-1})^{-1}$
How to select $\Pi$?



## Marginal Likelihood for Regularized Linear Regression



$Y = \Phi\theta + E,$ assume $E \in N(0, I)$

Parameterize prior covariance matrix.

$\theta \in N(0, \Pi(\alpha))$

That means

$Y \in N(0, \Phi\Pi(\alpha)\Phi^T + I)$

$\Rightarrow$
The Maximum likelihood (ML) estimate of $\alpha$ based on $Y, \Phi$ is

**Estimate of Regularization Matrix**

$$\hat{\alpha} = \arg\min Y^T Z(\alpha)^{-1} Y + \log\det Z(\alpha)$$
$$Z(\alpha) = \Phi\Pi(\alpha)\Phi^T + I$$

# ARX Model Priors



When estimating an ARX-model, we can think of the predictor

$$\hat{y}(t|\theta) = (1 - A(q))y(t) + B(q)u(t)$$

as made up of two impulse responses, $A$ and $B$. The vector $\theta$ should thus mimic two impulse responses, both typically exponentially decaying and smooth.We can thus have a reasonable prior for $\theta$:

$$P(\alpha_1, \alpha_2) = \begin{bmatrix} P^A(\alpha_1) & 0 \\ 0 & P^B(\alpha_2) \end{bmatrix} \qquad \text{Block Diagonal } A\&B$$

where the hyperparameters $\alpha$ describe decay and smoothness of the impulse responses.



# Typical Kernels



”Kernel”: Parameterization of $P$.
Several kernels exist: DC,TC,SS,... Recall: impulse response smooth and exponentially decaying

**DC kernel**

$E|b_k|^2 = C\lambda^k$, corr$(b_k, b_{k+1}) = \rho$
$P^B_{k,\ell} = C\lambda^{(k+\ell)/2}\rho^{|k-\ell|}; \quad \alpha = [C, \lambda, \rho]$

**TC kernel**

$E|b_k|^2 = C\lambda^k$, corr$(b_k, b_{k+1}) = \sqrt{\lambda}$
$P^B_{k,\ell} = C\min(\lambda^k, \lambda^\ell); \quad \alpha = [C, \lambda]$



# Software Issues



The MATLAB system Identification Toolbox, ver R2013b (released August 2013) now supports quadratic regularization for all linear and non-linear model estimation.

**The regularized criterion**

$$V_N(\theta) = \sum_{t=1}^N |\varepsilon(t,\theta)|^2 + \lambda(\theta - \theta^*)^T R(\theta - \theta^*),$$

is supported by a field `Regularization` in all the estimationOptions (`arxOptions, ssestOptions, procestOptions`) etc.:

```
opt.Regularization.Lambda
opt.Regularization.R
opt.Regularization.Nominal
```
($\theta^*$)

ARX-regularization tuning:

```
[L,R]=arxRegul(data,[na,nb,nk],Kernel)
```

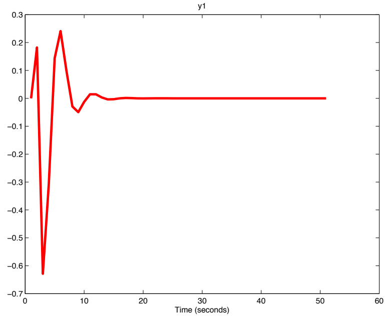


# Our Test Data: State-of-the-Art



Recall: The state-of-the art approach: Estimate SS models of different orders. Determine the order by the AIC criterion.

```
 for k=1:30
   m{k}= ssest(z,k);
end
(dum,n) = min(aic(:});
mss = m{n};
impulse(mss)
```

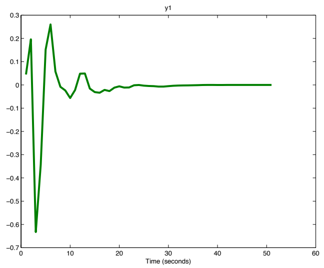


## Estimate a Model: Regularized ARX



Now, let us try an ARX model with na=5, nb=60. Estimate a regularization matrix with the 'TC' kernel (2 parameters, $C, \lambda$ each for the A and B parts):

```
 aopt = arxOptions;
(L,R) = arxRegul(z,[5 60 0],'TC');
aopt.Regularization.R = R;
aopt.Regularization.Lambda = L;
mr = arx(z,[5 60 0],aopt);
impulse(mr)
```



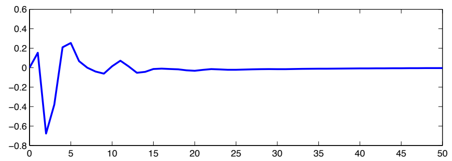



## The Oracle

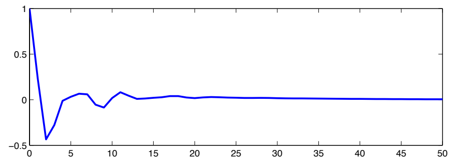


The examined data were obtained from a randomly generated model of order 30:

$$y(t) = G_0(q)u(t) + H_0(q)e(t)$$



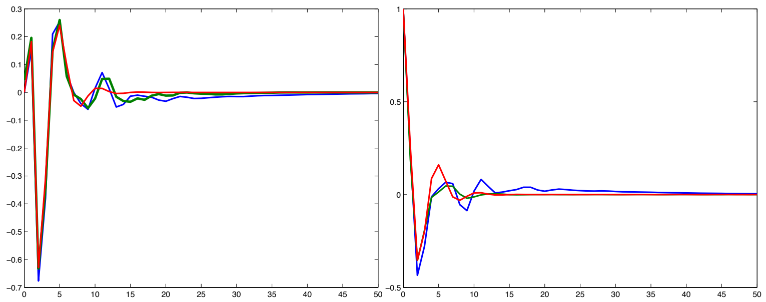
The input is Gaussian white noise with variance 1, and $e$ is white noise with variance 0.1. The impulse responses of $G$ and $H$ are shown at the right.





## How Well Did Our Models mss And mr Do?





$G$ : fit: **mss: 79.42% mr: 83.55%** $H$: fit **mss: 77.05%**, **mr: 81.59%**



## Surprise ?



ML beaten by an "outsider algorithm"!: That is a surprise!
There is a certain randomness in these data, but Monte-Carlo simulations substantiate the observed conclusion.

Even though ML is known to have the quoted optimal properties for best bias and variance, the observation is still not a contradiction.

Recall: Mean Square Error (MSE) = $|\text{Bias}|^2$ + Variance.

ML: Bias $\approx 0 \Rightarrow$: MSE = Variance = CR Lower bound for unbiased estimators

But with some bias, Variance could be clearly smaller then CRB

Recall for Lin Reg: CRB = $(\Phi\Phi^T)^{-1} > (\Phi\Phi^T + \Pi^{-1})^{-1}$ = MSE for best regularized estimated. More pronounced for short data

## Summary

- Frequency Domain Data and Spectral Analysis
  - Parametric models can be estimate just as well from FD data.
  - Non-parametric Spectral Analysis: Tune the window size for bias/variance trade-off
- IV methods
  - Good robust alternatives that do not require (or produce) a noise model
- Subspace Methods
  - Interesting alternative for MV linear systems. No iterative search for model. Theoretical properties worse. Several auxiliary variables to select
- Regularization
  - Useful complement to PEM methods, especially for short data records.