11. Prediction Error Methods (PEM)

- Description
- Optimal prediction
- Kalman filter
- Statistical results
- Computational aspects

Description

Determine the model parameter θ such that

$$e(t,\theta) = y(t) - \hat{y}(t|t-1;\theta)$$

is small

- $\hat{y}(t|t-1;\theta)$ is a prediction of $y(t)$ given the data up to and including time $t-1$ and based on θ
- a general linear predictor can be expressed as

$$\hat{y}(t|t-1;\theta) = L(q^{-1};\theta)y(t) + M(q^{-1};\theta)u(t)$$

  where L and M must contain one pure delay, *i.e.*,

$$L(0;\theta) = 0, M(0;\theta) = 0$$

Elements of PEM

One has to make the following choices, in order to define the method

- **Choice of model structure:** the parametrization of $G(q^{-1};\theta), H(q^{-1};\theta)$ and $\Lambda(\theta)$ as a function of θ
- **Choice of predictor:** the choice of filters L, M once the model is specified
- **Choice of criterion:** define a scalar-valued function of $e(t,\theta)$ that will assess the performance of the predictor

The most common way is to let $\hat{y}(t|t-1;\theta)$ be the *optimal mean square predictor*

The filters are chosen such that the prediction error have as small variance as possible

Loss function

Let N be the number of data points. Define the sample covariance matrix

$$R(\theta) = \frac{1}{N} \sum_{t=1}^{N} e(t,\theta) e^*(t,\theta)$$

$R(\theta)$ is a positive semidefinite matrix

In many cases, $R(\theta)$ is positive definite (e.g., when N is large)

A **loss function**

$$f(R(\theta))$$

is a scalar-valued function defined on the set of p.d.f. matrices R

f must be *monotonically increasing*, $i.e.$, let $X \succ 0$ and for any $\Delta X \succeq 0$

$$f(X + \Delta X) \geq f(X)$$

Example 1 $f(X) = \mathbf{tr}(WX)$ where $W \succ 0$ is a weighting matrix

$$f(X + \Delta X) = \mathbf{tr}(WX) + \mathbf{tr}(W\Delta X) \geq f(X)$$

($\mathbf{tr}(W\Delta X) \geq 0$ because if $A \succeq 0, B \succeq 0$, then $\mathbf{tr}(AB) \geq 0$)

Example 2 $f(X) = \det X$

$$\begin{aligned} f(X + \Delta X) - f(X) &= \det(X^{1/2}(I + X^{-1/2}\Delta X X^{-1/2})X^{1/2}) - \det X \\ &= \det X[\det(I + X^{-1/2}\Delta X X^{-1/2}) - 1] \\ &= \det X \left[\prod_{k=1}^{n}(1 + \lambda_k(X^{-1/2}\Delta X X^{-1/2})) - 1\right] \geq 0 \end{aligned}$$

The last inequalty follows from $X^{-1/2}\Delta X X^{-1/2} \succeq 0$, so $\lambda_k \geq 0$ for all k

Both examples satisfy $f(X + \Delta X) = f(X) \iff \Delta X = 0$

Procedures in PEM

- Choose a model structure of the form

$$y(t) = G(q^{-1};\theta)u(t) + H(q^{-1};\theta)\nu(t), \quad \mathbf{E}\,\nu(t)\nu(t)^* = \Lambda(\theta)$$

- Choose a predictor of the form

$$\hat{y}(t|t-1;\theta) = L(q^{-1};\theta)y(t) + M(q^{-1};\theta)u(t)$$

- Select a criterion function $f(R(\theta))$
- Determine $\hat{\theta}$ that minimizes the loss function f

Example: Least-squares method as a PEM

Use linear regression in the dynamics of the form

$$A(q^{-1})y(t) = B(q^{-1})u(t) + \varepsilon(t)$$

We can write $y(t) = H(t)\theta + \varepsilon(t)$ where

$$H(t) = \begin{bmatrix} -y(t-1) & \dots & -y(t-p) & u(t-1) & \dots & u(t-r) \end{bmatrix}$$
$$\theta = \begin{bmatrix} a_1 & \dots a_p & b_1 & \dots & b_r \end{bmatrix}^T$$

$\hat{\theta}$ that minimizes $(1/N)\sum_{t=1}^{N} \varepsilon^2(t)$ will gives a prediction of $y(t)$:

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

Hence, the prediction is in the form of

$$\hat{y}(t) = L(q^{-1};\theta)y(t) + M(q^{-1};\theta)u(t)$$

where $L(q^{-1};\theta) = 1 - \hat{A}(q^{-1})$ and $M(q^{-1};\theta) = B(q^{-1})$

note that $L(0;\theta) = 0$ and $M(0;\theta) = 0$,

so $\hat{y}$ uses the data up to time $t-1$ as required

The loss function in this case is $\mathbf{tr}(R(\theta))$ (quadratic in the prediction error)

Example: Maximum Likelihood estimation as a PEM

Suppose the noise $\nu(t)$ in the following model is *Gaussian* distributed

$$y(t) = G(q^{-1})u(t) + H(q^{-1})\nu(t), \quad \mathbf{E}\,\nu(t)\nu(s)^* = \Lambda\delta_{t,s}$$

Again drop θ from G, H, Λ and the unknowns are Λ, θ

The conditional likelihood function of $y(t)$ (conditioning on the initial conditions) is

$$L(\Lambda, \theta) = \frac{1}{(2\pi)^{N\cdot\dim(y)/2}\det\Lambda^{N/2}}\exp -\frac{1}{2}\sum_{t=1}^{N} e^T(t,\theta)\Lambda^{-1}e(t,\theta)$$

Take logarithms and ignore the constant term

$$\log L(\Lambda, \theta) = -\frac{N}{2}\log\det\Lambda - \frac{1}{2}\sum_{t=1}^{N} e^T(t,\theta)\Lambda^{-1}e(t,\theta)$$

Define $R(\theta) = (1/N) \sum_{t=1}^{N} e(t,\theta) e(t,\theta)^T$

$$\log L(\Lambda, \theta) = \frac{N}{2} \left\{ \log \det \Lambda^{-1} - \mathbf{tr}(\Lambda^{-1} R(\theta)) \right\}$$

Setting the gradient w.r.t Λ^{-1} to zero gives

$$\Lambda = R(\theta)$$

and the maximum likelihood problem turns to be

$$\text{maximize} \quad - \log \det R(\theta)$$

can be interpreted as a PEM using $\det$ as a loss function

Optimal prediction

Consider the general linear model

$$y(t) = G(q^{-1})u(t) + H(q^{-1})\nu(t), \quad \mathbf{E}\,\nu(t)\nu(s)^* = \Lambda\delta_{t,s}$$

(we drop argument θ in G, H, Λ for notational convenience)

Assumptions:

- $G(0) = 0, H(0) = I$
- $H^{-1}(q^{-1})$ and $H^{-1}(q^{-1})G(q^{-1})$ are asymptotically stable
- $u(t)$ and $\nu(s)$ are uncorrelated for $t < s$

Rewrite $y(t)$ as

$$\begin{aligned} y(t) &= G(q^{-1})u(t) + [H(q^{-1}) - I]\nu(t) + \nu(t) \\ &= G(q^{-1})u(t) + [H(q^{-1}) - I]H^{-1}(q^{-1})[y(t) - G(q^{-1})u(t)] + \nu(t) \\ &= \left\{H^{-1}(q^{-1})G(q^{-1})u(t) + [I - H^{-1}(q^{-1})]y(t)\right\} + \nu(t) \\ &\triangleq z(t) + \nu(t) \end{aligned}$$

- $G(0) = 0$ and $H(0) = I$ imply $z(t)$ contains $u(s), y(s)$ up to time $t - 1$
- Hence, $z(t)$ and $\nu(t)$ are uncorrelated

Let $\hat{y}(t)$ be an arbitrary predictor of $y(t)$

$$\begin{aligned} \mathbf{E}[y(t) - \hat{y}(t)][y(t) - \hat{y}(t)]^* &= \mathbf{E}[z(t) + \nu(t) - \hat{y}(t)][z(t) + \nu(t) - \hat{y}(t)]^* \\ &= \mathbf{E}[z(t) - \hat{y}(t)][z(t) - \hat{y}(t)]^* + \Lambda \geq \Lambda \end{aligned}$$

This gives a lower bound, Λ on the prediction error variance

The optimal predictor minimizes the prediction error variance

Therefore, $\hat{y}(t) = z(t)$ and is given by

$$\hat{y}(t|t-1) = H^{-1}(q^{-1})G(q^{-1})u(t) + [I - H^{-1}(q^{-1})]y(t)$$

The corresponding prediction error can be written as

$$e(t) = y(t) - \hat{y}(t|t-1) = \nu(t) = H^{-1}(q^{-1})[y(t) - G(q^{-1})u(t)]$$

- From $G(0) = 0$ and $H(0) = I$, $\hat{y}(t)$ depends on past data up to time $t-1$
- These expressions suggest asymptotical stability assumptions in $H^{-1}G$ and H^{-1}

Optimal predictor for an ARMAX model

Consider the model

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

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

For this particular case,

$$G(q^{-1}) = \frac{bq^{-1}}{1+aq^{-1}}, \quad H(q^{-1}) = \frac{1+cq^{-1}}{1+aq^{-1}}$$

Then the optimal predictor is given by

$$\hat{y}(t|t-1) = \frac{bq^{-1}}{1+cq^{-1}}u(t) + \frac{(c-a)q^{-1}}{1+cq^{-1}}y(t)$$

For computation, we use the recursion equation

$$\hat{y}(t|t-1) + c\hat{y}(t-1|t-2) = (c-a)y(t-1) + bu(t-1)$$

The prediction error is

$$e(t) = \frac{1+aq^{-1}}{1+cq^{-1}}y(t) - \frac{b}{1+cq^{-1}}u(t)$$

and it obeys

$$e(t) + ce(t-1) = y(t) + ay(t-1) - bu(t-1)$$

- The recursion equation requires an initial value, $i.e.$, $e(0)$
- Setting $e(0) = 0$ is equivalent to $\hat{y}(0|-1) = 0$
- The transient is not significant for large t

Kalman Filter

For systems given in a state-space form

$$x(t+1) = Ax(t) + Bu(t) + \nu(t)$$
$$y(t) = Cx(t) + \eta(t)$$

where $\nu(t), \eta(t)$ are mutually uncorrelated white noise with zero means and covariances R_1, R_2 resp.

The optimal one-step predictor of $y(t)$ is given by the *Kalman filter*

$$\hat{x}(t+1) = A\hat{x}(t) + Bu(t) + K[y(t) - C\hat{x}(t)]$$
$$\hat{y}(t) = C\hat{x}(t)$$

where K is the *steady-state Kalman gain*

The Kalman gain is given by

$$K = APC^*(CPC^* + R_2)^{-1}$$

and P is the solution to the *algebraic Riccati equation*:

$$P = APA^* + R_1 - APC^*(CPC^* + R_2)^{-1}CPA^*$$

- The predictor is *mean square optimal* if the disturbances are *Gaussian*
- For other distributions, the predictor is the *optimal linear predictor*

Example: Kalman filter of ARMAX model

Consider the model

$$y(t) + ay(t-1) = bu(t-1) + \zeta(t) + c\zeta(t-1)$$

where $|c| < 1$ and $\zeta(t)$ is zero mean white noise with variance λ^2

This model can be written in state-space form as

$$x(t+1) = \begin{bmatrix} -a & 1 \\ 0 & 0 \end{bmatrix} x(t) + \begin{bmatrix} b \\ 0 \end{bmatrix} u(t) + \begin{bmatrix} 1 \\ c \end{bmatrix} \zeta(t+1)$$
$$y(t) = \begin{bmatrix} 1 & 0 \end{bmatrix} x(t)$$

with $\nu(t) \triangleq \begin{bmatrix} 1 \\ c \end{bmatrix} \zeta(t+1)$ and then $R_1 = \lambda^2 \begin{bmatrix} 1 & c \\ c & c^2 \end{bmatrix}, R_2 = 0$

Since the last row of A is entirely zero, we can verify that P has the form

$$P = \lambda^2 \begin{bmatrix} 1+\alpha & c \\ c & c^2 \end{bmatrix}$$

where α satisfies

$$\alpha = (c-a)^2 + a^2\alpha - \frac{(c-a-a\alpha)^2}{1+\alpha}$$

There are two solutions, $\alpha = 0$ and $\alpha = c^2 - 1$

Hence, we pick $\alpha = 0$ to make P positive definite

The Kalman gain is therefore

$$K = \begin{bmatrix} -a & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & c \\ c & c^2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & c \\ c & c^2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right)^{-1} = \begin{bmatrix} c-a \\ 0 \end{bmatrix}$$

The one-step optimal predictor of the output is

$$\begin{aligned}\hat{x}(t+1) &= \begin{bmatrix}-a & 1\\ 0 & 0\end{bmatrix}\hat{x}(t) + \begin{bmatrix}b\\ 0\end{bmatrix}u(t) + \begin{bmatrix}c-a\\ 0\end{bmatrix}\left(y(t) - \begin{bmatrix}1 & 0\end{bmatrix}\hat{x}(t)\right)\\ &= \begin{bmatrix}-c & 1\\ 0 & 0\end{bmatrix}\hat{x}(t) + \begin{bmatrix}b\\ 0\end{bmatrix}u(t) + \begin{bmatrix}c-a\\ 0\end{bmatrix}y(t)\\ \hat{y}(t) &= \begin{bmatrix}1 & 0\end{bmatrix}\hat{x}(t)\end{aligned}$$

Then it follows that

$$\begin{aligned}\hat{y}(t) &= \begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}q+c & -1\\ 0 & q\end{bmatrix}^{-1}\begin{bmatrix}bu(t)+(c-a)y(t)\\ 0\end{bmatrix}\\ &= \frac{1}{q+c}[bu(t)+(c-a)y(t)]\\ &= \frac{bq^{-1}}{1+cq^{-1}}u(t) + \frac{(c-a)q^{-1}}{1+cq^{-1}}y(t)\end{aligned}$$

same result as in page 11-14

Theoretical result

Assumptions:

1. The data $\{u(t), y(t)\}$ are stationary processes

2. The input is persistently exciting

3. The Hessian $\nabla^2 f$ is nonsingular locally around the minimum points of $f(\theta)$

4. The filters $G(q^{-1}), H(q^{-1})$ are differentiable functions of θ

Under these assumptions, the PEM estimate is *consistent*

$$\hat{\theta} \to \theta, \quad \text{as} \quad N \to \infty$$

Statistical efficiency

For Gaussian disturbances the PEM method is *statistically efficient* if

- SISO: $f(\theta) = \mathbf{tr}(R(\theta))$
- MIMO:
  - $f(\theta) = \mathbf{tr}(WR(\theta))$ and $W = \Lambda^{-1}$ (the true covariance of noise)
  - $f(\theta) = \det(R(\theta))$

Computational aspects

I. Analytical solution exists

If the predictor is a linear function of the parameter

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

and the criterion function $f(\theta)$ is simple enough, *i.e.*,

$$f(\theta) = \mathbf{tr}(R(\theta)) = \frac{1}{N}\sum_{t=1}^{N} e(t,\theta)^2 = \frac{1}{N}\sum_{t=1}^{N}(y(t) - H(t)\theta)^2$$

It is clear that PEM is equivalent to the LS method

This holds for ARX or FIR models (but not for ARMAX and Output error models)

II. No analytical solution exists

It involves a nonlinear optimization for

- general criterion functions
- predictors that depend nonlinearly on the data

Example of numerical algorithms: Newton-Ralphson, Gradient based methods, Grid search

Typical issues in nonlinear minimization:

- solutions may consist of many local minima
- convergence rate and computational cost
- choice of initialization

Numerical example

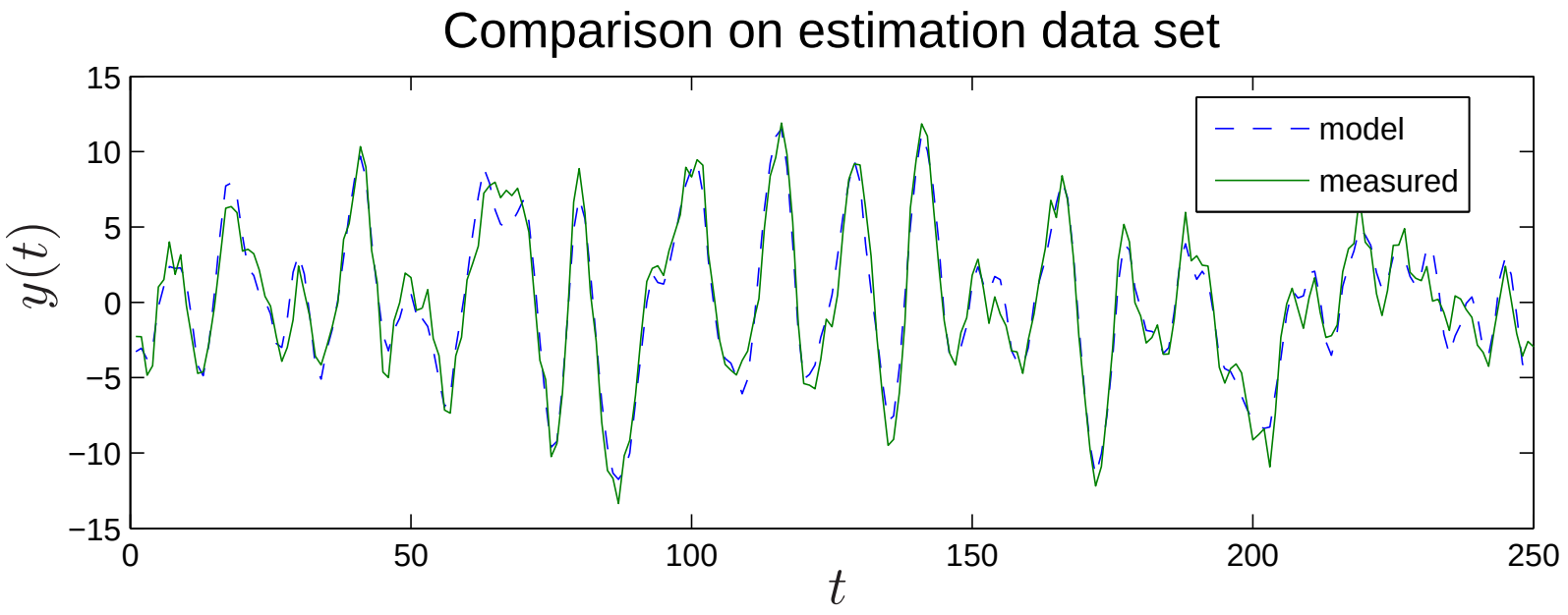
The true system is given by

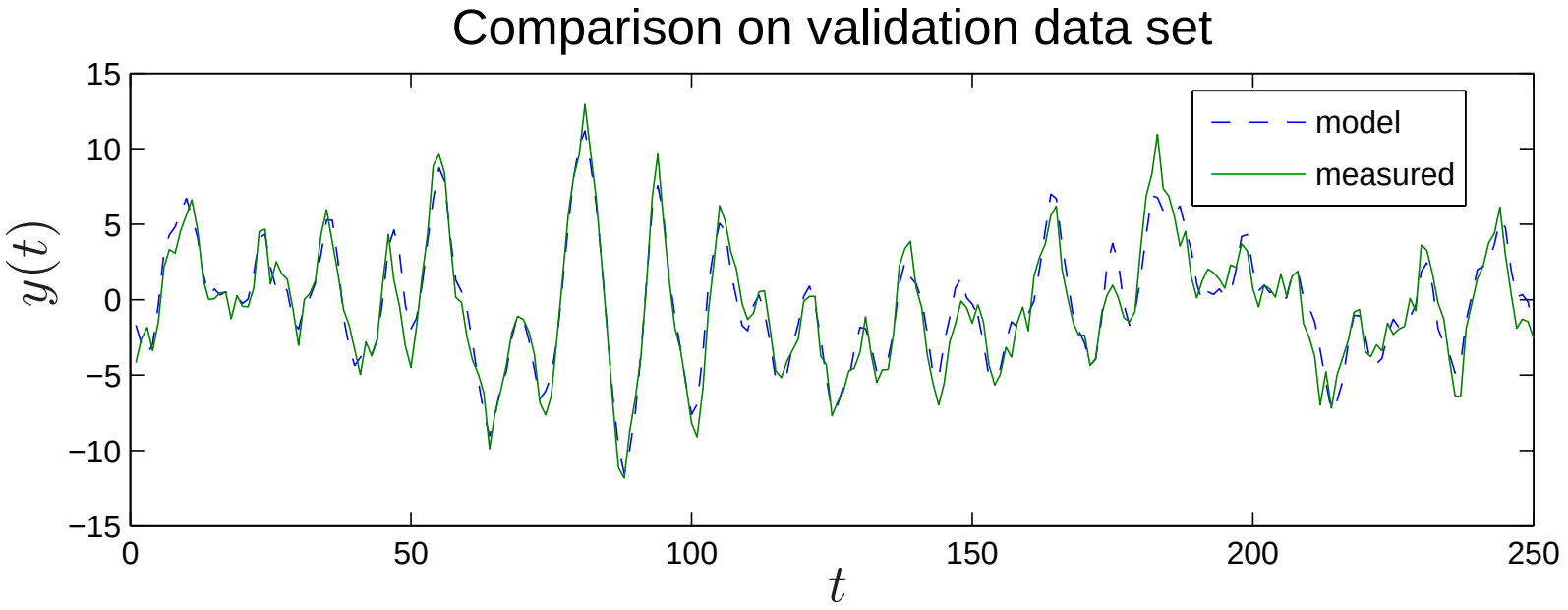
$$(1-1.5q^{-1}+0.7q^{-2})y(t) = (1.0q^{-1}+0.5q^{-2})u(t)+(1-1.0q^{-1}+0.2q^{-2})e(t)$$

- ARMAX model
- $u(t)$ is binary white noise, independent of $e(t)$
- $e(t)$ is white noise withzero mean and variance 1
- $N = 250$ (number of data points)

estimation

- assume the model structure and model order are known
- use `armax` command in MATLAB


Comparison on estimation data set
model
measured
y(t)
t
15
10
5
0
−5
−10
−15
0
50
100
150
200
250


Comparison on validation data set
model
measured
y(t)
t
15
10
5
0
−5
−10
−15
0
50
100
150
200
250

Example of MATLAB codes

```
%% Generate the data
N = 250; Ts = 1; u_var = 1; noise_var = 1;
a = [1 -1.5 0.7]; b = [0 1 .5]; c = [1 -1 0.2];
u = sign(randn(2*N,1))*sqrt(u_var); e = randn(2*N,1);
M = idpoly(a,b,c,1,1,noise_var,Ts);
y = sim(M,[u e]);
uv = u(N+1:end); ev = e(N+1:end); yv = y(N+1:end);
u = u(1:N); e = e(1:N); y = y(1:N);
DATe = iddata(y,u,Ts); DATv = iddata(yv,uv,Ts);

%% Identification
na = 2; nb = 2; nc = 2;
theta_pem = armax(DATe,[na nb nc 1]);   % ARMAX using PEM

%% Compare the measured output and the model output
[yhat1,fit1] = compare(DATe,theta_pem);
[yhat2,fit2] = compare(DATv,theta_pem);
```

```
t = 1:N;
figure;
subplot(2,1,1);plot(t,yhat1{1}.y,'--',t,y);
legend('model','measured');
title('Comparison on estimation data set','FontSize',16);
ylabel('y');xlabel('t');
subplot(2,1,2);plot(t,yhat2{1}.y,'--',t,yv);legend('y2','y');
legend('model','measured');
title('Comparison on validation data set','FontSize',16);
ylabel('y');xlabel('t');
```

References

Chapter 7 in
T. Söderström and P. Stoica, *System Identification*, Prentice Hall, 1989

Lecture on
Prediction Error Methods, System Identification (1TT875), Uppsala University,
`http://www.it.uu.se/edu/course/homepage/systemid/vt05`