10. Prediction Error Methods (PEM)

- description
- optimal prediction
- examples
- statistical results
- computational aspects

Description

idea: determine the model parameter θ such that

$$\varepsilon(t,\theta) = y(t) - \hat{y}(t|t-1;\theta) \qquad \text{is small}$$

$\hat{y}(t|t-1;\theta)$ is a prediction of $y(t)$ given the data up to time $t-1$ and based on θ

general linear predictor:

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

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

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

example: $\hat{y}(t|t-1;\theta) = 0.5y(t-1) + 0.1y(t-2) + 2u(t-1)$

Elements of PEM

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

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

we commonly consider the **optimal mean square predictor**

the filters N and M are chosen such that the prediction error has small variance

Loss function

let N be the number of data points

sample covariance matrix:

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

$R(\theta)$ is a positive semidefinite matrix (and typically pdf when N is large)

loss function: scalar-valued function defined on positive matrices R

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

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

1. choose a model structure of the form

$$y(t) = G(L;\theta)u(t) + H(L;\theta)e(t), \quad \mathbf{E}e(t)e(t)^T = \Lambda(\theta)$$

2. choose a predictor of the form

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

3. select a criterion function $V(\theta) := f(R(\theta))$
4. determine $\hat{\theta}$ that minimizes the loss function V

   (some time we use V_N to emphasize that V depends on the sample size N)

Least-squares method as a PEM

use linear regression in the dynamics of the form

$$A(L)y(t) = B(L)u(t) + e(t)$$

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

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

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

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

hence, the prediction is in the form of

$$\hat{y}(t) = N(L;\theta)y(t) + M(L;\theta)u(t)$$

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

note that $N(0;\theta) = 0$ 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)

Optimal prediction

consider the general linear model

$$y(t) = G(L;\theta)u(t) + H(L;\theta)e(t), \quad \mathbf{E}[e(t)e(s)^T] = \Lambda\delta_{t,s}$$

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

assumptions:

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

rewrite $y(t)$ as

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

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

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

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

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 the **optimal predictor** is given by

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

the corresponding **optimal prediction error** can be written as

$$\begin{aligned} \varepsilon(t) &= y(t) - \hat{y}(t|t-1) = e(t) \\ &= H^{-1}(L)[y(t) - G(L)u(t)] \end{aligned}$$

- 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

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

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

for this particular case,

$$G(L) = \frac{bL}{1+aL}, \quad H(L) = \frac{1+cL}{1+aL}$$

then the optimal predictor is given by

$$\hat{y}(t|t-1) = \left(\frac{bL}{1+cL}\right) u(t) + \left(\frac{(c-a)L}{1+cL}\right) y(t)$$

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

$$\varepsilon(t) = \left(\frac{1+aL}{1+cL}\right) y(t) - \left(\frac{bL}{1+cL}\right) u(t)$$

and it obeys

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

- the recursion equation requires an initial value, *i.e.*, $\varepsilon(0)$
- setting $\varepsilon(0) = 0$ is equivalent to $\hat{y}(0|-1) = y(0)$
- the transient is not significant for large t
- to find $\hat{\theta}_{\mathrm{pem}}$, we minimize $V(\theta)$ over (a, b, c) (nonlinear optimization)

Loss function minimization

PEM estimate $\hat{\theta}$ minimizes

$$V(\theta) := f(R(\theta)) = f\left(\frac{1}{N}\sum_{t=1}^{N} \varepsilon(t,\theta)\varepsilon(t,\theta)^T\right)$$

to find a local minimizer using numerical methods, it requires

$$\frac{\partial V}{\partial \theta} = \frac{\partial f}{\partial R} \cdot \frac{1}{N}\sum_{t=1}^{N} \frac{\partial}{\partial \theta}\left[\varepsilon(t,\theta)\varepsilon(t,\theta)^T\right]$$

example: scalar system and using $f(R) = \mathbf{tr}(R)$ will give

$$V(\theta) = (1/N)\sum_{t=1}^{N} \varepsilon(t,\theta)^2, \quad \nabla V(\theta) = (2/N)\sum_{t=1}^{N} \varepsilon(t,\theta)\nabla_\theta \varepsilon(t,\theta)$$

and $\nabla_\theta \varepsilon$ is typically nonlinear in θ

example: $\nabla_\theta \varepsilon(t,\theta)$ for ARMA(1,1) (special case of page 10-13)

$$
\begin{aligned}
\varepsilon(t) &= \left(\frac{1+aL}{1+cL}\right) y(t) \\
\frac{\partial \varepsilon(t)}{\partial a} &= \left(\frac{L}{1+cL}\right) y(t) \\
\frac{\partial \varepsilon(t)}{\partial c} &= -\frac{(1+aL)L}{(1+cL)^2} y(t) = -\frac{L}{(1+cL)} \varepsilon(t,\theta)
\end{aligned}
$$

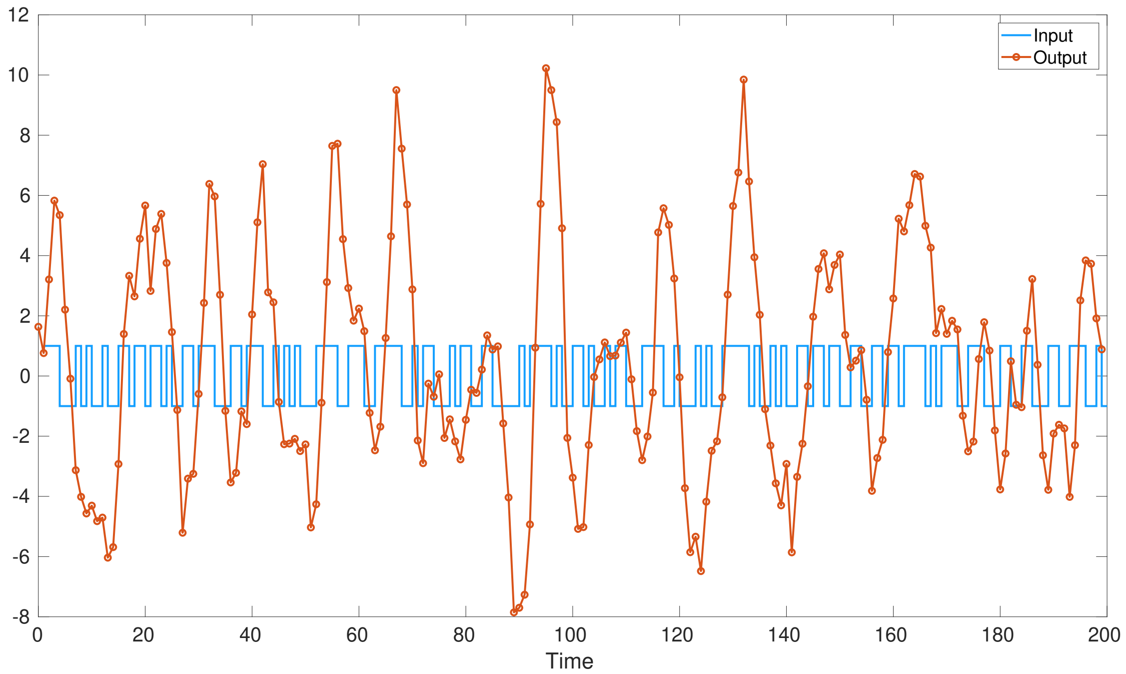
input arguments of `pem` command in system identification toolbox:

- input/output $\{(u_i, y_i)\}_{i=1}^N$
- initial parameter: $\theta^{(0)}$ for the search method in optimization
- imposing constraint of θ (if any)

Numerical example

the true system (dgp) is ARMAX(2,2,2)

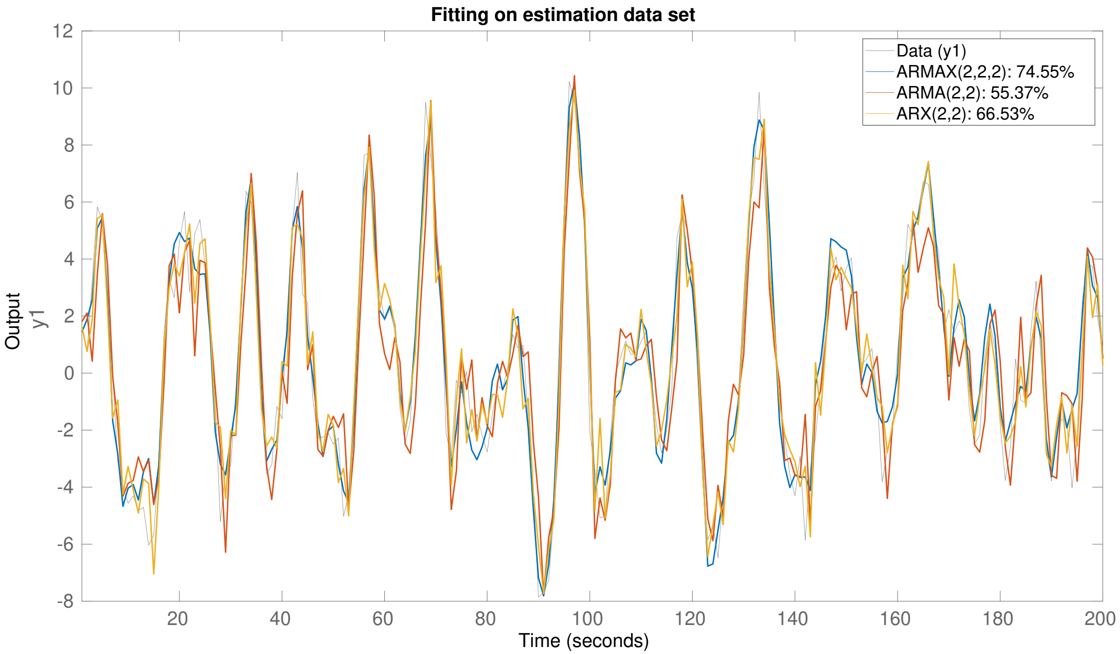
$$(1 - 1.5L + 0.7L^2)y(t) = (1.0L + 0.5L^2)u(t) + (1 - 1.0L + 0.2L^2)e(t)$$



both u, e are white with unit variance; u is binary and independent of e

estimation: `armax` and `arx` commands to estimate three models

- ARMAX(2,2,2): guess the model structure correctly
- ARMA(2,2): make no use of input in estimation
- ARX(2,2): no consideration in the noise dynamics



using a simpler model (ARX) or neglecting u yielded worse result than using the model with correct structure

Example of MATLAB codes

```
% Generate the data
N = 200; Ts = 1; t = (0:Ts:Ts*(N-1))'; noise_var = 1;
a = [1 -1.5 0.7]; b = [0 1 .5]; c = [1 -1 0.2];
u = idinput(N,'PRBS');
e = sqrt(noise_var)*randn(N,1);
dgp = idpoly(a,b,c,1,1,noise_var,Ts); % data generating process
opt = simOptions('AddNoise',true,'NoiseData',e);
y = sim(dgp,u,opt); DAT = iddata(y,u,Ts);

% Identification
m = armax(DAT,[2 2 2 1]); % [na nb nc nk] ARMAX(2,2,2)
m1 = armax(DAT,[2 0 2 1]); % ARMA(2,2)
m2 = arx(DAT,[2 2 1]); % ARX(2,2) uses the LS method

% Compare the measured output and the model output
compare(DAT,m,m1,m2,1) ; % Use '1' to compare the 1-step prediction
```

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 $f(R)$ is simple enough, *i.e.*,

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

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

numerical algorithms: Newton-Ralphson, Gradient based methods

typical issues in nonlinear minimization:

- problem has many local minima
- convergence rate and computational cost
- choice of initialization

Feasible set of parameters

suppose the ground-truth system is described by

$$\mathcal{S}: \quad y(t) = G_0(L)u(t) + H_0(L)e(t), \quad E[e(t)e(\tau)^T] = \Lambda_0 \delta_{t,\tau}$$

and that we assume the model $\mathcal{M}(\theta)$ in estimation process

consider all model parameters that make the model matched with the true system

$$D(\mathcal{M}) = \{\theta \mid G_0(L) = G(L;\theta),\ H_0(L) = H(L;\theta),\ \Lambda_0 = \Lambda(\theta)\}$$

we denote the set of all feasible parameters as $D(\mathcal{M})$

all three possibilities of $D(\mathcal{M})$: empty set, unique member, many members

Properties of PEM estimate

properties of PEM estimate depends on

- existence of members in $D(\mathcal{M})$
- choice of loss function

$$V_N(\theta) := f(R(\theta)) = f\left(\frac{1}{N}\sum_{t=1}^{N}\varepsilon(t,\theta)\varepsilon(t,\theta)^T\right)$$

$\hat{\theta}_N$ minimizes $V_N(\theta)$ where N data samples are used

we examine consistency of $\hat{\theta}_N$ (when $N \to \infty$)

Consistency property

assumptions:

1. the data $\{u(t), y(t)\}$ are quasi-stationary processes
2. the input is persistently exciting
3. $\nabla V_N(\theta)$ and $\nabla^2 V_N(\theta)$ are continuous; $\nabla^2 V_N(\theta)$ is non-singular in neighbors of local minima
4. both G and H are differentiable functions of θ and uniformly stable
5. $D(\mathcal{M})$ is not empty

under these assumptions, the PEM estimate is **consistent**

$$\hat{\theta}_N \xrightarrow{p} \theta^*, \quad \text{as} \quad N \to \infty$$

Statistical efficiency

assumption: $D(\mathcal{M})$ contains only one member, θ^*

- define $s(t) = \nabla_\theta \varepsilon(t, \theta^\star)$ and $F = \left.\frac{\partial f}{\partial R}\right|_{R=\Lambda}$
- PEM estimate has a limiting normal distribution

$$\sqrt{N}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, P)$$

$$P = \left(\mathbf{E}[s(t)Fs(t)]\right)^{-1} \mathbf{E}[s(t)F\Lambda Fs(t)^T] \left(\mathbf{E}[s(t)Fs(t)]\right)^{-1}$$

where $P \succeq (\mathbf{E}[s(t)\Lambda^{-1}s(t)^T])^{-1}$ (covariance has a lower bound)

- P achieves its lower bound (PEM is efficient) in each of the following cases:
  - y is scalar and $f(R) = \mathbf{tr}(R)$
  - $f(R) = \mathbf{tr}(WR)$ and choose $W = \Lambda^{-1}$ (inverse of noise covariance)
  - $f(R) = \log\det(R)$

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`