

# System Identification, Lecture 5

Kristiaan Pelckmans (IT/UU, 2338)

Course code: 1RT880, Report code: 61800 - Spring 2012  
F, FRI Uppsala University, Information Technology

14 Februari 2012

# Exam(ple)

- Understand what written in lecture notes/slides.
- Open book.
- Insight.
  - Is the PEM equal to BLUE estimate?
  - Is the covariance of the LS estimate for an  $ARX(d)$  system equal to the covariance of the PEM estimate for this system?
  - Do we need to define explicitly a stochastic setup for PEM to make sense?
- Exercises.
  - What is the one-step ahead optimal predictor corresponding to a  $MA(2)$  timeseries model?
  - Give the maximum likelihood estimator of an  $AR(2)$  model with Gaussian zero mean, unit variance white noise. When is this estimate efficient?

- Give the Steepest Descent Method for an ARMAX(1,1) model.

# Lecture 5

- Prediction Error Method.
- Optimal Prediction.
- Computational Methods.
- Analytical Results.
- Implementations.

# Prediction Error Method

General:

- Why identification?
- Task: Best possible prediction.
- Predict as most accurately as possible.
- Find parameter  $\theta$  which leads to best predictions.



DILBERT: © Scott Adams/Dist. by United Feature Syndicate, Inc

Model:

$$y_t = g(\mathbf{x}_t; \theta_0) + e_t$$

Estimating the parameter  $\theta$ ?

LS:

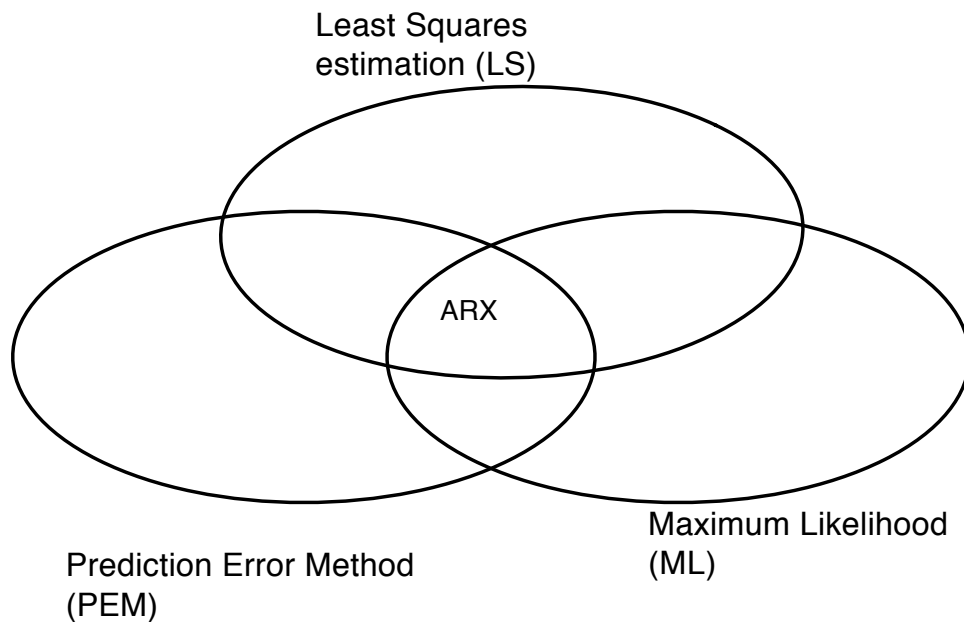
$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{2} \sum_{t=1}^n \left( y_t - g(\mathbf{x}_t; \theta) \right)^2$$

ML: Assume a stochastic model (PDF) as  $\{Y_1, \dots, Y_n\} \sim f(\theta_0)$ , then

$$\theta_n = \underset{\theta}{\operatorname{argmax}} \ln L(Y_1, \dots, Y_n; \theta)$$

PEM: Where  $\hat{y}_t(\theta, \mathbf{x}_t, y_{t-1}, \dots, y_1)$  is best prediction of  $y_t$  based on the model,  $\mathbf{x}_t$  and all past observations  $y_1, \dots, y_t$

$$\theta_n^p = \underset{\theta}{\operatorname{argmin}} \frac{1}{2} \sum_{t=1}^{n-1} \left( y_{t+1} - \hat{y}_{t+1}(\theta, \mathbf{x}_{t+1}, y_t, \dots, y_1) \right)^2$$



### Difference LS-PEM:

- LS unbiased if noise  $\{e_t\}$  independent of  $\{X_t\}$
- First approach: instrumental variable.
- More principled: PEM.

# Optimal Predictor

- If  $\{e_1, \dots, e_{t-1}, e_t\}$  realizations of zero mean white noise random variable, what is than best predictor of  $e_t$  based on  $\{e_1, \dots, e_{t-1}\}$ ?

- If system

$$y_t = b_1 u_{t-1} + \dots + b_d u_{t-d} + e_t$$

with  $\{e_t\}_t$  zero mean white noise. Given  $\{\dots, u_{t-1}, u_t\}$  and  $\{\dots, y_t\}$ , what is the best prediction of  $y_{t+1}$ ?

- If timeseries system

$$y_t - (a_1 y_{t-1} + \dots + a_d y_{t-d}) = e_t$$

with  $\{e_t\}_t$  zero mean white noise. Given  $\{\dots, y_t\}$ , what is the best prediction of  $y_{t+1}$ ?



- If timeseries model  $MA(d)$

$$y_t = (1 + c_1q^{-1} + \dots + c_dq^{-d})e_t = C(q^{-1}; \theta_0)e_t$$

with  $\{e_t\}_t$  zero mean white noise. Given  $\{\dots, y_t\}$ , what is the best prediction of  $y_{t+1}$ ?

- Hence at time  $t + 1$  one has

$$y_{t+1} = (C(q^{-1}; \theta_0) - 1) e_{t+1} + e_{t+1}$$

And  $e_t = C^{-1}(q^{-1}; \theta_0)y_t$ . Thus

$$y_{t+1} = (C(q^{-1}; \theta_0) - 1) C^{-1}(q^{-1}; \theta_0)y_{t+1} + e_{t+1}$$

Which is equal to

$$y_{t+1} = (1 - C^{-1}(q^{-1}; \theta_0) - 1) y_{t+1} + e_{t+1}$$

Since  $C^{-1}$  is monic too, the best predictor is

$$\hat{y}_{t+1|t} = (1 - C^{-1}(q^{-1}; \theta_0)) y_{t+1}$$

- In general model

$$y_{t+1} = H(q^{-1}, \theta_0)u_{t+1} + G(q^{-1}, \theta_0)e_{t+1}$$

where  $H(q^{-1}, \theta_0) = 1 + h_1q^{-1} + \dots$  and  $G(q^{-1}, \theta_0) = 1 + g_1q^{-1} + \dots$ .

- Rewrite as optimal predictor

$$\hat{y}_{t|t-1} = L_1(q^{-1}, \theta_0)u_t + L_2(q^{-1}, \theta_0)y_t$$

where  $L_2(1, \cdot) = 0$ .

- How:

$$y_{t+1} = (\dots)u_{t+1} + (\dots)e_t + e_{t+1}$$

and

$$e_t = (\dots)u_t + (\dots)y_t$$

Then

$$y_{t+1} = L_1u_{t+1} + L_2y_t + e_{t+1}$$

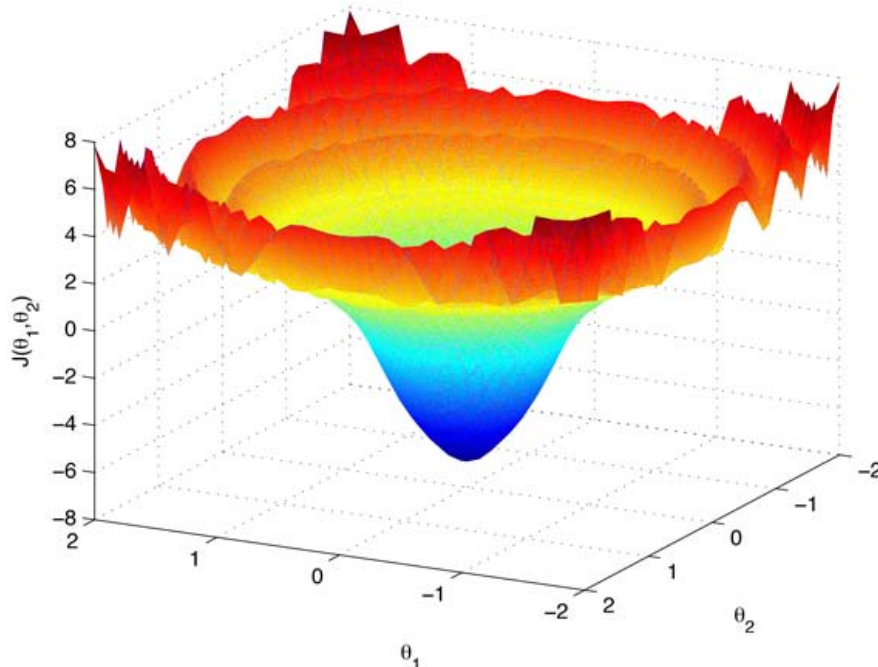
- Optimal predictor

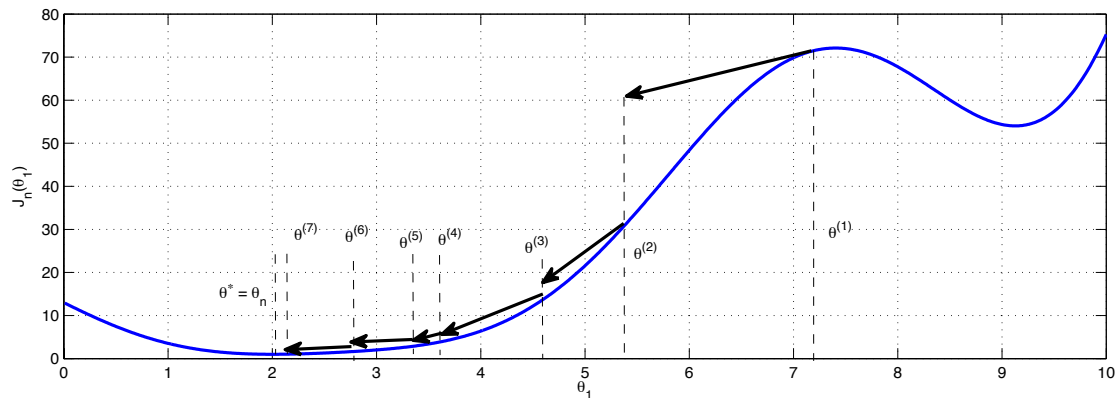
$$\hat{y}_{t+1|t} = (G^{-1}(q^{-1}, \theta_0)H(q^{-1}, \theta_0))u_{t+1} + (1 - G^{-1}(q^{-1}, \theta_0))y_t$$

# Computational Methods

$$\theta_n^p = \theta^* = \operatorname{argmin}_{\theta} J_n(\theta) = \frac{1}{2} \sum_{t=1}^n (y_t - \hat{y}_t(\theta, \dots))^2$$

- In general not rewritable as LIP.
- Need for global optimization tools.





- Numerical  $\Rightarrow$  Iterative Procedure
- Algorithm generating sequence  $\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \dots\}$  with parameters getting closer to the minimizer  $\theta^{(i)} \rightarrow \theta^* = \theta_n$ .
- Choice of starting value  $\theta^{(0)}$ .

- Update rule

$$\theta^{(i+1)} \leftarrow \theta^{(i)} + \boxed{\text{step}}$$

- Such that

$$J_n(\theta^{(0)}) \geq J_n(\theta^{(1)}) \geq \dots \geq J_n(\theta^{(m)}) \approx J_n(\theta^*)$$

- Guarantee: speed and convergence.

## Common choices

- Steepest Descent

$$\theta^{(i+1)} \leftarrow \theta^{(i)} - \gamma_i \nabla J_n(\theta^{(i)})$$

- Newton-Raphson:

$$\theta^{(i+1)} \leftarrow \theta^{(i)} - \gamma_i [\nabla^2 J_n(\theta^{(i)})]^{-1} \nabla J_n(\theta^{(i)})$$

- Gauss-Newton:

$$\theta^{(i+1)} \leftarrow \theta^{(i)} - \gamma_i \mathbb{E}[\nabla^2 J_n(\theta^{(i)})]^{-1} \nabla J_n(\theta^{(i)})$$

- Quasi-Newton:

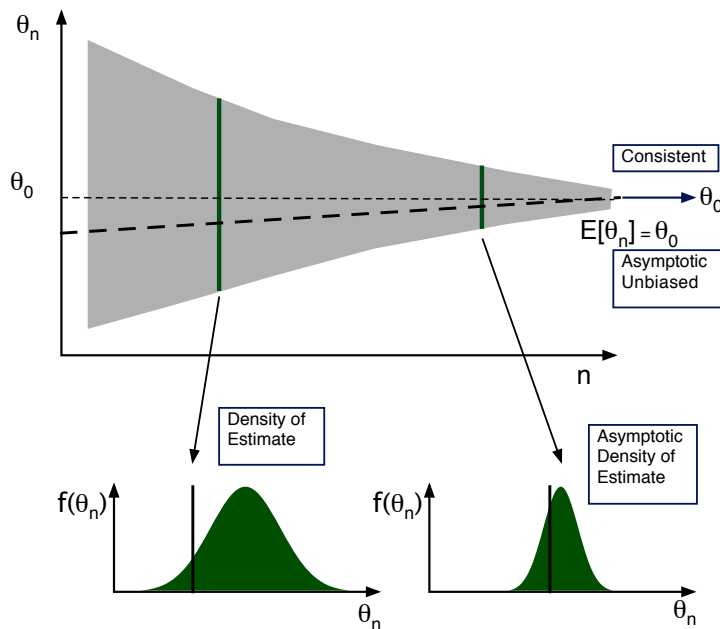
$$\theta^{(i+1)} \leftarrow \theta^{(i)} - \gamma_i H(\theta^{(i)}) \nabla J_n(\theta^{(i)})$$

with  $H(\theta^{(i)}) \approx [\nabla^2 J_n(\theta^{(i)})]^{-1}$

## Issues

- Gradient  $\nabla J_n(\theta^{(i)})$  of Prediction error  $J_n$  with respect to parameters  $\theta$ .
- Hessian  $\nabla^2 J_n(\theta^{(i)})$  and invert: cumbersome
- Generalized Least Squares (GLS): practical approach

# Analytical Results



Properties of an estimator  $\theta_n$ :

- Consistent.
- Biased for finite  $n$ .
- Asymptotic Unbiased.
- Asymptotic Sample Distribution.



## Proper assumptions:

- Stochastic assumption on signals.
- Identifiable at optimum.
- Bounded Gradient.
- Input Data is rich enough.

## Limit distribution estimates:

$$\sqrt{n}(\theta_n^p - \theta_0) \sim \mathcal{N}(0, P)$$

for  $n \rightarrow \infty$ , and  $P$  Fisher Information matrix. Achieves asymptotically Cramér-Rao lowerbound!

$$P = \frac{\sigma^2}{n} \mathbb{E} [\psi_t \psi_t^T]^{-1}, \quad \psi_t = \left. \frac{d\epsilon_t(\theta)}{d\theta} \right|_{\theta=\theta_0}$$

# Conclusions

- PEM, ML, LS.
- Model vs. Predictor.
- Computational Approach.
- Asymptotic Results.
- Practice.