System Identification, Lecture 5

Kristiaan Pelckmans (IT/UU, 2338)

Course code: 1RT880, Report code: 61800 - Spring 2015
F, FRI Uppsala University, Information Technology

17 April 2015

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 θ 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 θ?

LS:

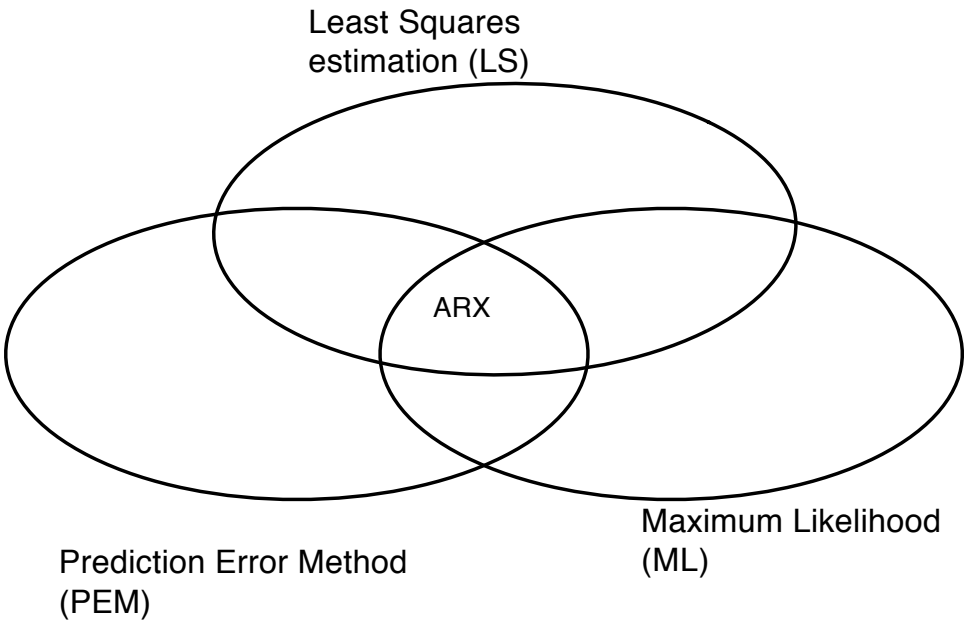
$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{2} \sum_{t=1}^{n} \Big(y_t - g(\mathbf{x}_t; \theta) \Big)^2$$

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

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

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

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



Difference LS-PEM:

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

Optimal Predictor

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

- If system

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

with $\{e_t\}_t$ zero mean white noise. Given $\{\ldots, u_{t-1}, u_t\}$, $\{b_1, \ldots, b_d\}$ and $\{\ldots, y_t\}$, what is the best prediction of y_{t+1}?

- If timeseries system

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

with $\{e_t\}_t$ zero mean white noise. Given $\{\ldots, y_t\}$, , $\{a_1, \ldots, a_d\}$. What is the best prediction of y_{t+1}?

- If timeseries model MA(d)

$$y_t = (1 + c_1 q^{-1} + \cdots + c_d q^{-d}) e_t = C(q^{-1}; \theta_0) e_t$$

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

- Hence at time $t+1$ one has

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

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

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

Which is equal to

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

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

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

- In general model

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

where $H(q^{-1}, \theta_0) = 1 + h_1 q^{-1} + \dots$ and $G(q^{-1}, \theta_0) = 1 + g_1 q^{-1} + \dots$.

- 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:

$$y_{t+1} = (...)u_{t+1} + (...)e_{t+1} + e_{t+1}$$

and

$$e_t = (...)u_t + (...)y_t$$

Then

$$y_{t+1} = L_1 u_{t+1} + L_2 y_t + e_{t+1}$$

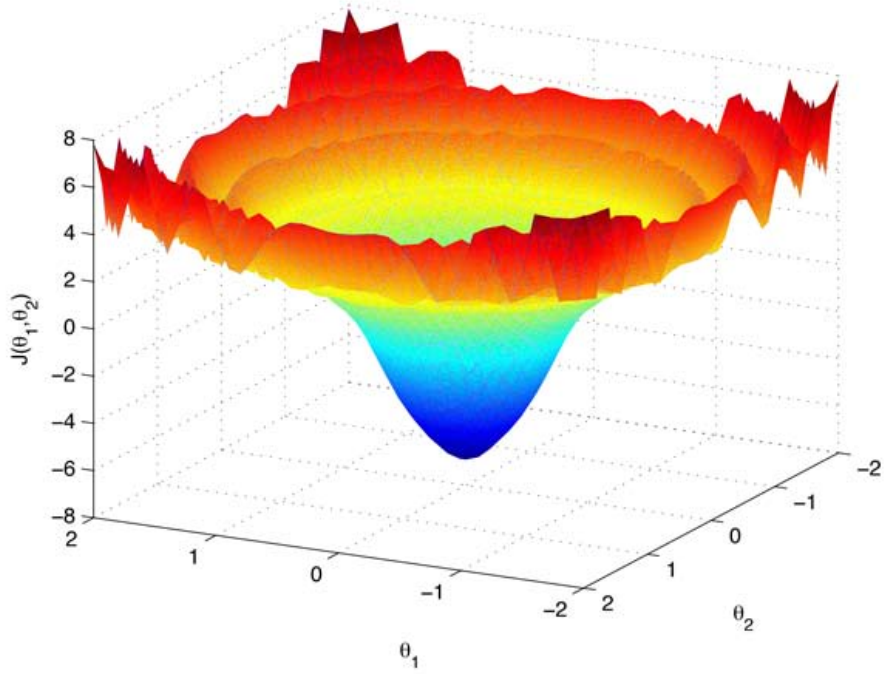
- Optimal predictor

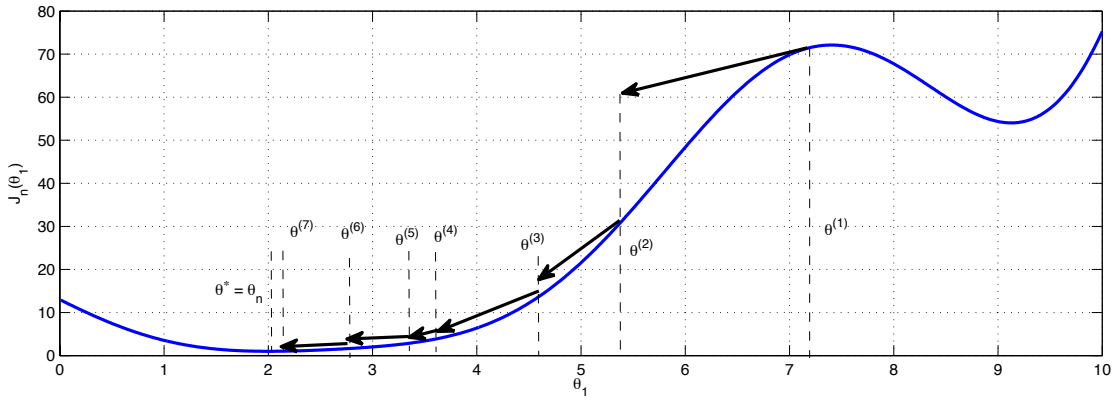
$$\hat{y}_{t+1|t} = \big(G^{-1}(q^{-1}, \theta_0)H(q^{-1}, \theta_0)\big)u_{t+1} + \big(1 - G^{-1}(q^{-1}, \theta_0)\big)y_t$$

Computational Methods

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

- 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

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

- Guarantee: speed and convergence.

Common choices

- Steepest Descent

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

- Newton-Raphson:

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

- Guass-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:

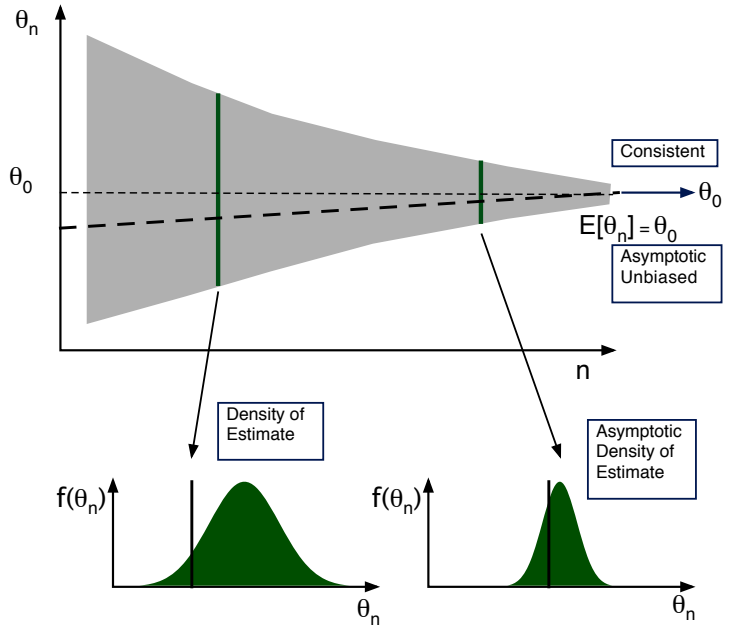
$$\theta^{(i+1)} \leftarrow \theta^{(i)} - \gamma_i H(\theta^{(i)}) \nabla J_n(\theta^{(i)})$$

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 θ.
- Hessian $\nabla^2 J_n(\theta^{(i)})$ and invert: cumbersome
- Generalized Least Squares (GLS): practical approach

Analytical Results



Properties of an estimator θ_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 $n \to \infty$, and P Fisher Information matrix. Achieves asymptotically Cramér-Rao lowerbound!

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

Conclusions

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