System Identification, Lecture 4

Kristiaan Pelckmans (IT/UU, 2338)

Course code: 1RT875, Report code: 61806,
F, FRI Uppsala University, Information Technology

26 January 2010

Lecture 4

- Prediction Error Methods (PEM) (Ch. 7)

The Least Squares Method

- Chapter 4: the least squares method applied to static (deterministic) linear regression models ($\varphi(t)$ deterministic).

- What happens when we consider dynamic models?

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

Write as

$$y(t) = \varphi^T(t)\theta + e(t)$$

where

$$\varphi(t) = (-y(t-1), \ldots, -y(t-n_a), u(t-1), \ldots, u(t-n_b))^T$$

and

$$\theta = \left(a_1, \ldots, a_{n_a}, b_1, \ldots, b_{n_b}\right)^T$$

- Least Squares estimator:

$$\hat{\theta}_{LS} = \left(\frac{1}{N} \sum_{t=1}^{N} \varphi(t)\varphi^T(t) \right)^{-1} \frac{1}{N} \sum_{t=1}^{N} \varphi(t)y(t)$$

Properties: Assume the 'true' system can be described as

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

Then, the least squares estimate $\hat{\theta}_{LS}$ will be consistent ($\hat{\theta}_{LS} \rightarrow \theta_0$ as $N \rightarrow \infty$), if

- $E[\varphi(t)\varphi^T(t)]$ is nonsingular.

- $E[\varphi(t)v(t)] = 0$

The first condition will be satisfied in most cases. A few exceptions

- The input is not persistently exiting of order n_b.
- The data is noise-free ($v(t) \equiv 0$) and the model order is chosen too high (this implies that A and B have common factors).
- The system operates under feedback with a low order regulator.

The second condition is in most cases *not* satisfied. A notable exception is when $e(t)$ is white noise.

Modifications of the Least Squares Methods

The second constraint is relaxed as follows:

- Prediction error Methods. Models the noise as well!
- The Instrumental variable methods (IV methods) - modify the normal equations of the least-squares estimator.

Prediction Error Methods (PEM)

Idea:

- Models the noise as well $\rightarrow$ stochastic model, i.e. the outputs of the models are not deterministic.
- Minimize the prediction errors $\epsilon(t,\theta) = y(t) - \hat{y}(t|t-1,\theta)$
- The LS estimator is a special case, where

$$\epsilon(t,\theta) = y(t) - y(t|t-1,\theta) = y(t) - \varphi^T(t)\theta.$$

Hence, a general methodology applicable to a wide range of model structures.

Examples.

Find the optimal predictor, $\hat{y}(t|t-1)$ for the following systems assuming $E[e(t)] = 0$ and $E[e(t)e(s)] = \lambda^2 \delta_{ts}$ (notice that $y(t|t-1)$ is a function of $\{(u(s), y(s))\}_{s<t}$).

- $y(t) = e(t)$
- $(1 - 0.1q^{-1})y(t) = -0.5q^{-1}u(t) + e(t)$
- $(1 - 0.1q^{-1})y(t) = -0.5q^{-1}u(t) + (1 - 0.8q^{-1})e(t)$

Predictions

A predictor can be described as a filter that predicts the output of a dynamic system given past measured input- and output signals. Design the predictor as

- Choose the model structure $\mathcal{M}$, e.g. ARX, OE or ARMAX
- Choose the predictor $\hat{y}(t|t-1,\theta)$. A general predictor can be viewed as

$$y(t|t-1,\theta) = L_1(q^{-1},\theta)y(t) + L_2(q^{-1},\theta)u(t)$$

  where L_1 and L_2 are such that they only take past measurements into account.

Optimal Predictor

We will here consider the general model structure:

$$y(t) = G(q^{-1}, \theta)u(t) + H(q^{-1}, \theta)e(t)$$

where $E[e(t)] = 0$ and $E[e(t)e(s)] = \lambda^2 \delta_{ts}$.

Goal: Find the optimal mean least square predictor $\hat{y}(t, t - 1, \theta)$, i.e. solve

$$\min_{y(t|t-1,\theta)} E[\epsilon(t)\epsilon^T(t)]$$

where $\epsilon(t) = y(t) - y(t|t-1, \theta)$ is the prediction error and $\hat{y}(t|t-1, \theta)$ depends on the past measurements only.

Results: Under the assumptions that

- $z(t)$ only depends on past measurements.
- $u(t)$ and $e(s)$ are uncorrelated for $t < s$

then

$$\hat{y}(t|t-1,\theta) = H^{-1}(q^{-1},\theta)G^{-1}(q^{-1},\theta)u(t) + \left(I - H^{-1}(q^{-1},\theta)\right) y(t) \quad (1)$$

is the optimal mean square predictor, and $e(t)$ the prediction error, and

$$\begin{aligned} \epsilon(t,\theta) &= y(t) - \hat{y}(t|t-1,\theta) \\ &= H^{-1}(q^{-1},\theta)y(t) - G^{-1}(q^{-1},\theta)u(t) \\ &= e(t). \end{aligned}$$

Hence

$$E[\epsilon(t,\theta)\epsilon^T(t,\theta)] = \Lambda(\theta)$$

Optimal Prediction for State Space Models

As an alternative to the model structure:

$$y(t) = G(q^{-1}, \theta)u(t) + H(q^{-1}, \theta)e(t),$$

it is common to use a state-space model with states $(x(t))_t \subset \mathbb{R}^n$

$$\begin{cases} x(t+1) = F(\theta)x(t) + B(\theta)u(t) + v(t) \\ y(t) = C(\theta)x(t) + e(t) \end{cases}$$

where $v(t)$ and $e(t)$ are uncorrelated white noise sequences with zero mean and covariance matrices $\mathbf{R}_1$ and $\mathbf{R}_2$.

In this case the optimal mean square predictor is given by the **Kalman filter** (see p.196).

Cost Function

How do we find the best model in the model structure?

- Minimize the prediction errors $\epsilon(t, \theta)$ for all t. How?
- Choose a criterion function $V_N(\theta)$ to minimize

$$\hat{\theta} = \operatorname*{argmin}_{\theta} V_N(\theta)$$

where $V_N(\theta)$ depends on $\epsilon(t, \theta)$ is a suitable manner.

Depending on the choice of model structures, predictor filters and criterion function, the minimization of the loss function is simple/difficult.

For MISO systems the following criterion function is most often used:

$$V_N(\theta) = \frac{1}{N}\sum_{t=1}^{N} \epsilon^2(t,\theta).$$

In general, the cost function is choses as

$$V_N(\theta) = h(\mathbf{R}_N(\theta)),$$

where $h : \mathbb{R}^{n\times n} \to \mathbb{R}$ is a scalar-valued, monotonically increasing function, and $\mathbf{R}_N(\theta)$ is the covariance matrix of the prediction errors, or

$$\mathbf{R}_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \epsilon(t,\theta)\epsilon^T(t,\theta).$$

Ex. $h(\cdot) = \mathrm{tr}(\cdot)$ or $h(\cdot) = \det(\cdot)$.

A PEM Algorithm

In order to make a PEM, the user has to make the following choices:

- Choice of model structures. How should G^{-1}, H^{-1} and Λ be parametrized by θ?
- Choice of predictor. Usually the optimal mean square predictor is used.
- Choice of criterion function $V_N(\theta)$. A scalar-valued function of all prediction errors $\{\epsilon(t, \theta)\}_t$ which will assess the performance of the predictor used.

Computational Aspects

Analytical (closed-form) solutions exists If the predictor is 'linear-in-the-parameters', or

$$\hat{y}(t|t-1,\theta) = \varphi^T(t)\theta,$$

and the associate criterion V_N is simple enough, a closed form solution may exists. For example if

$$V_n(\theta) = \frac{1}{N}\sum_{t=1}^{N}\epsilon^2(t,\theta),$$

PEM is equivalent to OLS. This holds for example for ARX or FIR models, but *not* for ARMAX or OE models.

No Analytical (closed-form) solutions exists In general criteria, and for predictors that are not 'linear-in-the-parameters', a numerical search algorithm is required to find θ that minimizes $V_N(\theta)$.

Numerical minimization

- Nonlinear optimization $\rightarrow$ local minima may exist.
- Time-consuming (convergence rate) and computationally complex.
- Initialization.

Different (standard) methods available:

- The **Newton-Raphson** algorithm.

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} - \alpha_k \left(V_N''(\hat{\theta}^{(k)})^{-1} V'(\hat{\theta}^{(k)}) \right)$$

  The gradient (Hessian) of the cost-function are often computationally expensive to calculate. Fast Convergence.

- The **Gauss-Newton** algorithm is a computationally less demanding algorithm, with a (theoretically) slower rate of convergence.

- **Gradient-based** methods are simple to apply, but even slower convergence rates.

- **Grid-search** Search the whole parameter space. VERY time-consuming.

Theoretical Analysis

Assumptions

- The signals $(u(t), y(t))_t$ are stationary stochastic processes.
- The input sequence if PE.
- $V_N''(\theta)$ is nonsingular around the minimum points of $V_N(\theta)$.
- The filters $G^{-1}(q^{-1}, \theta)$ and $H^{-1}(q^{-1}, \theta)$ are smooth differentiable functions of the parameter vector.

What happens with the estimate $\hat{\theta}_N$ as $N \to \infty$?

Consistency:

$$\begin{cases} \hat{\theta}_\infty \triangleq \lim_{N\to\infty} \hat{\theta}_N = \operatorname{argmin}_\theta V_\infty(\theta) \\ \operatorname{argmin}_\theta V_\infty(\theta) = \lim_{N\to\infty} \operatorname{argmin}_\theta \frac{1}{2} \sum_{t=1}^{N} \epsilon^2(t,\theta) \approx E[\epsilon^2(t,\theta)] \end{cases}$$

The PEM estimates are robust and efficient:

- As $N \to \infty$, $\hat{\theta}_N$ converges to a minimum point of V_∞
- If the model class includes the 'true' system $\mathcal{S}$, then the PEM is SI ($\hat{\theta}_\infty \in \mathcal{D}_T$)
- If $\mathcal{S}$ is PI, then the PEM is consistent (or $\hat{\theta}_N \to \theta_0$ as $N \to \infty$).

Asymptotic Distributions: Asymptotic distributions of the parameter estimates (assuming that the model in PI), or $\hat{\theta}_N \to \theta_0$.

- The parameter estimate errors are asymptotically Gaussian distributed, with zero mean and variance $\mathbf{P}$,

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \to \mathcal{N}(0_n, \mathbf{P})$$

- for SISO systems, the covariance matrix $\mathbf{P}$ is given as

$$\mathbf{P} = \Lambda E\left[\psi(t, \theta_0)\psi^T(t, \theta_0)\right]^{-1}$$

where

$$\psi(t, \theta_0) = -\frac{\partial \epsilon(t, \theta)}{\partial \theta}$$

and $\Lambda = E[e(t)e^T(t)]$.

Accuracy of linear regression for static/dynamic case:

Static case

- $\hat{\theta}_N$ unbiased
- Asymptotically Gaussian

$$\mathbf{P} = \Lambda \left(\frac{1}{N} \sum_{t=1}^{N} \varphi(t, \theta_0) \varphi^T(t, \theta_0) \right)^{-1}$$

Dynamic case $(N \to \infty)$

- $\hat{\theta}_N$ consistent
- Asymptotically Gaussian as $\mathcal{N}(0_n, \mathbf{P})$ with

$$\mathbf{P} = \Lambda E \left[\varphi(t, \theta_0) \varphi^T(t, \theta_0) \right]^{-1}$$

Statistical Efficiency:

- A method is said to be statistically efficient if its estimates have the smallest possible variance.
- The smallest possible variance of any (asymptotically) unbiased estimator is given by the Cramér-Rao lowerbound.
- For Gaussian disturbances, the PEM is statistically efficient. (equivalent to the Maximum Likelihood estimator) if
  - Single-output: $V_N(\theta) = \frac{1}{N}\sum_{t=1}^{N} \epsilon^2(t,\theta)$.
  - Multiple-output:

$$V_N(\theta) = \mathrm{tr}\left(\mathbf{S}\mathbf{R}_N(\theta)\right),$$

    where $\mathbf{S} = \Lambda^{-1}(\theta_0)$,
  - or $V_N(\theta) = \det(\mathbf{R}_N(\theta))$

Approximation

The true system is often more complex than the model structure (under-parametrization, $\mathcal{D}_T$ is empty)

- Still $\hat{\theta}_N$ converges to a minimum point of $V_N(\theta)$ as $N \to \infty$.
- We cannot expect $G(q^{-1}, \theta) \equiv G_0(q^{-1})$ or $H(q^{-1}, \theta) \equiv H_0(q^{-1})$.
- The model-fit can be controlled by pre-filtering the data,

$$u_F(t) = F(q^{-1})u(t), \quad y_F(t) = F(q^{-1})y(t)$$

  or by choosing an appropriate input.
- The OE model structure is useful.

Conclusions

- The PEM is a general method to obtain a parametric model of a dynamic system. The following choices define a prediction error method:
  - Choice of model structure.
  - Choice of predictor.
  - Choice of criterion function.
- The PEM principle is to minimize the prediction errors given a certain model structure and predictor.
- The PEM principle leads to parameter estimates that have several nice properties (in general consistent and statistically efficient estimates).