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 ($\boldsymbol{\varphi}(t)$ deterministic).
- What happens when we consider dynamic models?

$$A(q^{-1},\boldsymbol{\theta})y(t) = B(q^{-1},\boldsymbol{\theta})u(t) + e(t)$$
$$\Rightarrow \qquad y(t) = \boldsymbol{\varphi}^T(t)\boldsymbol{\theta} + e(t)$$

where

$$\boldsymbol{\varphi}(t) = [\,-y(t-1)\ \ldots\ -y(t-n_a)\ u(t-1)\ \ldots\ u(t-n_b)]^T$$
$$\boldsymbol{\theta} = [a_1\ \ldots\ a_{n_a}\ b_1\ \ldots\ b_{n_b}]^T$$

Properties of the least squares estimate

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

Properties: Assume that the true system can be described as

$$y(t) = \boldsymbol{\varphi}^T(t)\boldsymbol{\theta}_0 + v(t)$$

Results: The estimate $\hat{\boldsymbol{\theta}}_{LS}$ will be consistent ($\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0$ as $N \to \infty$) if

(i) $E\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^T(t)$ is nonsingular.

(ii) $E\boldsymbol{\varphi}(t)v(t) = \mathbf{0}$.

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 (which implies that $A(q^{-1})$ and $B(q^{-1})$ 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 Method

To relax the second constraint, we will in the following examine two different ways to modify the least-squares method:

(i) Prediction error methods. Model the noise as well!

(ii) The instrumental variables methods (IV methods) – modifying the normal equations associated with the least-squares estimate.

Prediction Error Methods (PEM)

Idea:

- Model the noise as well $\Rightarrow$ stochastic models, *i.e.*, the outputs from the models are not deterministic.
- Minimize the prediction errors $\varepsilon(t, \boldsymbol{\theta}) = y(t) - \hat{y}(t|t-1, \boldsymbol{\theta})$. The least-squares method is a special case of this approach; consider the prediction error

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

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 $Ee(t) = 0$, $Ee(t)e(s) = \delta_{s,t}\lambda^2$.

Notice that $\hat{y}(t|t-1)$ is a function of $\{y(s), u(s)\}_{s=-\infty}^{t-1}$.

a) $y(t) = e(t)$

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

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

Predictions

A predictor can be described as a filter that predicts the output of a dynamic system given old measured outputs and inputs. Design the predictor by

(i) Choosing the model structure of $y(t)$, *e.g.*, ARX, OE, or ARMAX.

(ii) Choosing the predictor, $\hat{y}(t|t-1, \boldsymbol{\theta})$. A general predictor can be viewed as

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

where $L_1(q^{-1}, \boldsymbol{\theta})$ and $L_2(q^{-1}, \boldsymbol{\theta})$ are constrained such that $\hat{y}(t|t-1, \boldsymbol{\theta})$ depends on past data.

Optimal Prediction

We will here consider the general model structure

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

where $E[e(t)e^T(s)] = \boldsymbol{\Lambda}(\boldsymbol{\theta})\delta_{t,s}$ and $G(0, \boldsymbol{\theta}) = 0$.

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

$$\min_{\hat{y}(t|t-1)} E\varepsilon(t)\varepsilon^T(t)$$

where $\varepsilon(t) = y(t) - \hat{y}(t|t-1)$ is the prediction error, and $\hat{y}(t|t-1)$ depends on $\{y(s), u(s)\}_{s=-\infty}^{t-1}$.

Results:

Under the assumptions that

(i) $z(t)$ only depends on past measurements

(ii) $u(t)$ and $e(s)$ are uncorrelated for $t < s$

then

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

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

$$\begin{aligned} \varepsilon(t, \boldsymbol{\theta}) &= y(t) - \hat{y}(t|t-1, \boldsymbol{\theta}) \\ &= H^{-1}(q^{-1}, \boldsymbol{\theta})\left[y(t) - G(q^{-1}, \boldsymbol{\theta})u(t)\right] \\ &= e(t) \end{aligned}$$

Hence,

$$E\varepsilon(t, \boldsymbol{\theta})\varepsilon^T(t, \boldsymbol{\theta}) = \boldsymbol{\Lambda}(\boldsymbol{\theta})$$

Optimal Prediction for State Space Models

As an alternative to the model structure:

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

it is often common to use state-space models:

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

where $v(t)$ and $e(t)$ are uncorrelated white noise sequences with zero mean and covariance matrices $R_1(\boldsymbol{\theta})$ and $R_2(\boldsymbol{\theta})$.

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

Cost Function

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

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

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} V_N(\boldsymbol{\theta})$$

where $V_N(\boldsymbol{\theta})$ depends on $\varepsilon(t, \boldsymbol{\theta})$ is a suitable manner.

Depending on the choice of model structure, predictor filters and criterion function, the minimization of the loss function is more or less difficult.

For single-output systems the following criterion function is most often used

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

In general, the cost function is chosen as

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

where $h(\cdot)$ is a scalar-valued monotonically increasing function, and $\mathbf{R}_N(\boldsymbol{\theta})$ is the sample covariance matrix of the prediction errors,

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

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

A PEM Algorithm

To define a PEM the user has to make the following choices:

- Choice of model structure. How should $G(q^{-1},\boldsymbol{\theta})$, $H(q^{-1},\boldsymbol{\theta})$ and $\Lambda(\boldsymbol{\theta})$ be parameterized?
- Choice of predictor $\hat{y}(t|t-1,\boldsymbol{\theta})$. Usually the optimal mean square predictor is used.
- Choice of criterion function $V(\boldsymbol{\theta})$. A scalar-valued function of all the prediction errors $\varepsilon(1,\boldsymbol{\theta}),\ldots,\varepsilon(N,\boldsymbol{\theta})$, which will assess the performance of the predictor used.

Computational Aspects

I. Analytical solution exists

If the predictor is a linear function of the unknown parameters,

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

and the criterion function $V_N(\boldsymbol{\theta})$ is simple enough, a closed form solution can be found. For example, when

$$V_N(\boldsymbol{\theta}) = \frac{1}{N}\sum_{t=1}^{N}\varepsilon^2(t,\boldsymbol{\theta}) = \frac{1}{N}\sum_{t=1}^{N}\left(y(t)-\boldsymbol{\varphi}^T(t)\boldsymbol{\theta}\right)^2,$$

it is clear that the PEM is equivalent to linear regression (the least squares method). This holds for example for ARX or FIR models but **not** for ARMAX and OE models.

II. No analytical solution exists

For general criterion functions, and predictors that depend non-linearly on the data, a numerical search algorithm is required to find the $\boldsymbol{\theta}$ that minimizes $V_N(\boldsymbol{\theta})$.

Numerical minimization:

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

Different (standard) methods available:

- The **Newton-Raphson** algorithm

$$\hat{\boldsymbol{\theta}}^{(k+1)} = \hat{\boldsymbol{\theta}}^{(k)} - \alpha_k [V''(\hat{\boldsymbol{\theta}}^{(k)})]^{-1} V'(\hat{\boldsymbol{\theta}}^{(k)})$$

  The derivatives of the loss function can be computationally complex to evaluate. Fast convergence.

- The **Gauss-Newton** algorithm is a computationally less intensive algorithm with a theoretically lower rate of convergence which can be used as an alternative.

- **Gradient based** methods are simpler to apply, but has a slow convergence rate.

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

Theoretical Analysis

Assumptions

- The data $\{u(t), y(t)\}$ are stationary processes.
- The input is persistently exciting.
- $V_N''(\boldsymbol{\theta})$ is nonsingular around the minimum points of $V_N(\boldsymbol{\theta})$.
- The filters $G(q^{-1}, \boldsymbol{\theta})$ and $H(q^{-1}, \boldsymbol{\theta})$ are smooth differentiable functions of the parameter vector.

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

Consistency:

$$\hat{\boldsymbol{\theta}}_\infty \triangleq \lim_{N\to\infty} \hat{\boldsymbol{\theta}}_N = \arg\min_{\boldsymbol{\theta}} V_\infty(\boldsymbol{\theta})$$

$$V_\infty(\boldsymbol{\theta}) = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \varepsilon^2(t, \boldsymbol{\theta}) = E\varepsilon^2(t, \boldsymbol{\theta})$$

The PEM estimates are robust and consistent:

- As $N \to \infty$, $\hat{\boldsymbol{\theta}}_N$ converges to a minimum point of $V_\infty(\boldsymbol{\theta})$.
- If the model structure includes the true system (D_T non-empty) then the PEM is system identifiable ($\hat{\boldsymbol{\theta}}_\infty \in D_T$).
- If there is a unique vector $\hat{\boldsymbol{\theta}}_\infty$ that gives an exact description of the system (D_T contains one point), then the system is parameter identifiable. The PEM estimate is **consistent** ($\hat{\boldsymbol{\theta}}_N \to \boldsymbol{\theta}_0$ as $N \to \infty$).

Asymptotic distribution: Asymptotic distribution of the parameter estimates (assuming that the model is parameter identifiable, $\hat{\boldsymbol{\theta}}_N \to \boldsymbol{\theta}_0$)

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

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \to N(0, \mathbf{P})$$

- For single-output systems the covariance matrix of the parameter estimates are given by

$$\mathbf{P} = \boldsymbol{\Lambda} \left[E\Psi(t, \boldsymbol{\theta}_0)\Psi^T(t, \boldsymbol{\theta}_0) \right]^{-1}$$

  where

$$\Psi(t, \boldsymbol{\theta}) = -\left(\frac{\partial \varepsilon(t, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T$$

  and $Ee(t)e^T(t) = \boldsymbol{\Lambda}$.

Accuracy of linear regression for static/dynamic case

Static case (N finite)

- $\hat{\boldsymbol{\theta}}$ unbiased.
- $\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)$ is Gaussian distributed $N(0, P)$,

$$P = \boldsymbol{\Lambda} \left(\frac{1}{N} \sum_{t=1}^{N} \boldsymbol{\varphi}(t, \boldsymbol{\theta}_0) \boldsymbol{\varphi}^T(t, \boldsymbol{\theta}_0) \right)^{-1}$$

Dynamic case (for $N \to \infty$).

- $\hat{\boldsymbol{\theta}}$ is consistent.
- $\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)$ is asymptotically Gaussian distributed $N(0, P)$.

$$P = \boldsymbol{\Lambda} \left[E \boldsymbol{\varphi}(t, \boldsymbol{\theta}_0) \boldsymbol{\varphi}^T(t, \boldsymbol{\theta}_0) \right]^{-1}$$

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 lower bound.
- For Gaussian disturbances the PEM method is statistically efficient (equivalent to the maximum likelihood (ML) method) if
  - Single-output: $V_N(\boldsymbol{\theta}) = \frac{1}{N} \sum_{t=1}^{N} \varepsilon^2(t, \theta)$.
  - Multi-output: $V_N(\boldsymbol{\theta}) = \operatorname{tr}(S R_N(\boldsymbol{\theta}))$ and $S = \Lambda^{-1}(\boldsymbol{\theta}_0)$, or $V_N(\boldsymbol{\theta}) = \det(R_N(\boldsymbol{\theta}))$.

Approximation

The true system is often more complex than the model structure (under-parametrization, D_T is empty).

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

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

  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).

- Approximation. The PEM is useful also for under-parameterized models. The model-fit can be controlled by pre-filtering the data, or by choosing an appropriate input.
- If the prediction errors depend linearly on the parameter vector the PEM estimates are obtained through linear regression (e.g., ARX and FIR models).
- In the case of more complicated model structures a nonlinear search algorithm is required to obtain the PEM estimates (e.g., ARMAX, OE, etc.).